LEARNING INTERACTION KERNELS AND EMERGENT BEHAVIORS FOR SECOND ORDER INTERACTING AGENT SYSTEMS

by

Jason Miller

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

June 2021

Abstract

Modeling the complex interactions of systems of particles or agents is a fundamental problem across the sciences, from physics and biology, to economics and social sciences. In this work, we consider second-order, heterogeneous, multivariable models of interacting agents or particles, within simple environments. We describe a nonparametric inference framework to efficiently estimate the interaction kernels which drive these dynamical systems. We develop a complete learning theory which establishes strong consistency and optimal nonparametric min-max rates of convergence for the estimators, as well as provably accurate predicted trajectories. The estimators exploit the structure of the equations in order to overcome the curse of dimensionality; furthermore we describe a fundamental coercivity condition which ensures that the interaction kernels can be learned and relates to the minimal singular value of the learning matrix. The numerical algorithm presented to build the estimators is parallelizable, performs well on high-dimensional problems, and its performance is tested on a variety of complex dynamical systems.

We are often interested in collective dynamical systems exhibiting emergent behaviors with complicated interaction kernels, and with kernels which are parameterized by a single unknown parameter. We provide extensive numerical evidence that the estimators provide faithful approximations to these interaction kernels, and provide accurate predictions for trajectories started at new initial conditions, both throughout the "training" time interval in which the observations were made, and much beyond.

We demonstrate these features on prototypical systems displaying collective behaviors, ranging from opinion dynamics, flocking dynamics, self-propelling particle dynamics, to synchronized oscillator dynamics. We also consider the problem of learning interaction kernels in these dynamical systems constrained to evolve on Riemannian manifolds. The models are based on interaction kernels depending on pairwise Riemannian distances between agents, with agents interacting locally along the direction of the shortest geodesic connecting them.

Lastly, we build accurate and predictive models of the underlying mechanisms of celestial motion. By modeling the major Astronomical Bodies in the Solar system as pairwise interacting bodies, we generate extremely accurate dynamics can provide a unified explanation to the observation data, especially in terms of reproducing the perihelion precession of Mars, Mercury, and the Moon.

Primary Reader: Mauro Maggioni

Secondary Reader: Fei Lu

Acknowledgements

I have been blessed with so many wonderful people in my life who have supported and encouraged me throughout my academic career, and without whom this thesis would never have been completed. First and foremost, I have to thank my remarkable parents, Gregory and Vibha Miller, who are my original supporters and have helped me in more ways than I could possibly describe. Their wisdom and belief in me was so critical throughout my undergraduate and graduate careers. My brother and sister, Devin and Priya Miller, were a wonderful sounding board and cheerleaders for me during the very focused periods necessary to do research. My experience throughout my Ph.D. has been the greatest and most fulfilling period of intellectual and personal growth in my life. The chief architect of that has been my advisor Mauro Maggioni. He gave me a constant supply of interesting problems to think about, taught me how to conceptualize and perform academic research, and provided nothing but support throughout my time at Johns Hopkins. He was also honest when my work wasn't up to par, when I should have considered a sharper argument or proof, or when my writing was lacking. This constructive feedback was critical and much appreciated as it greatly improved my research output and ability. I am also very grateful for the collaboration and mentorship that I received from Ming Zhong and Sui Tang. Our time working together on various research papers taught me a tremendous amount of mathematics and was integral to my learning, development, and this thesis. In addition to my immediate family and close research mentors, there are other people who played a

key role in this thesis and who I am grateful for. My fellow Ph.D. students here at Johns Hopkins, especially Joshua Agterberg, Zachary Pisano, Vittorio Loprinzo, Noah Wichrowski, Philip Kerger, and Yashil Sukurdeep, have been wonderful friends, sources of new ideas and ways of thinking, and the lifeblood of day-to-day living as I worked my way through my doctorate. I also give a special thanks to Pallavi Ravada, who was there in the beginning and helped encourage me in many ways throughout my mathematics journey. My friends, extended family, and former colleagues all also played a role and they are too numerous to list here, but they know who they are and have my thanks. A special thank you to the professors and mentors who helped support and guide me to take the leap into graduate school, Brian Parshall (University of Virginia), Michael Hill (UCLA), and Hilda Ochoa-Brillembourg. They each provided me with valuable mentorship and did so openly and kindly, for that and more I am grateful to them. Lastly, I want to thank the many professors here at Johns Hopkins who were excellent teachers, and helped to create a community and the feeling among the graduate students that you too can do meaningful research. Some of these professors include, James Fill, Carey Priebe, Fei Lu, John Wierman, Avanti Athreya, Donniell Fishkind, Yannis Kevrikidis, Daniel Robinson, Nicolas Charon, Daniel Naiman, Laurent Younes, and Marc Kamionkowski. My thesis represents a tremendous amount of work and is something I am very proud to have my name on but, in many ways, what I am most proud of are the relationships, the experiences, and new knowledge that are implicitly and explicitly represented by this document and that I will carry with me long after I leave Johns Hopkins – for this and much more I am grateful.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Physical, biological, and social systems across all scales of complexity and size can often be described as dynamical systems written in terms of interacting agents (e.g. particles, cells, humans, planets, ...). Rich theories have been developed to explain the collective behavior of these interacting agents across many fields including astronomy, particle physics, economics, social science, and biology. Examples include predator-prey systems, molecular dynamics, coupled harmonic oscillators, flocking birds or milling fish, human social interactions, and celestial mechanics, to name a few. In order to encompass many of these examples, we will consider a rather general family of second-order, heterogeneous (the agents can be of different types), interacting (the acceleration of an agent is a function of properties of the other agents) agent system that includes external forces, masses of the agents, multivariable interaction kernels, and an additional environment variable that is a dynamical property of the agent (for example, a firefly having its luminescence varying in time). We propose a learning approach that combines machine learning and dynamical systems in order to provide highly accurate dynamical models of the observation data from these systems. The model and learning framework presented in sections 2.2-2.4 includes a very large number of relevant systems and allows for their modeling. Clustering of opinions [76, 40, 16, 98]

is a simple first-order case that exhibits clustering. Flocking of birds [46, 43, 41] can be modeled as the behavior of a second-order system that exhibits an emergent shared velocity of all agents. Milling of fish [38, 1, 5, 37] may be modeled as a large-time behavior of a second-order system (in 2 or 3-dimensions), with a non-collective force from the environment. A model of oscillators (fireflies) that sync and swarm together, and have their dynamics governed by their positions and a phase variable ξ, was studied by [127, 102, 101, 100]. There are also models that include both energy and alignment interaction kernels, a particular case of this is the anticipation dynamics model from [121], which we also consider in this work. These dynamics exhibit a wide range of emergent behaviors, and as shown in [136, 128, 43, 62, 36, 10, 98], the behaviors can be studied when the governing equations are known. However, if the equations are not known and the data consists of only trajectories, we still wish to develop a model that can make accurate predictions of the trajectories and discover a dynamical form that accurately reflects their emergent properties. To achieve this, we present a provably optimal learning algorithm that is accurate, captures emergent behavior for large time, and, by exploiting the structure of the collective dynamical system, avoids the curse of dimensionality.

In the second chapter, we consider various numerical simulations that fit into the framework introduced in the first chapter and focus on whether or not our estimators can capture delicate emergent behaviors that these systems exhibit. Emergent behavior in collective dynamics, such as clustering of opinions [76, 40, 16, 98], flocking of birds [46, 43, 41], milling of fish [38, 1, 5, 37], and concentric trajectories of planetary motion [86], is among one of the most interesting phenomena in macroscopic and microscopic scale systems. It occurs in systems used across many disciplines, including biology, social science, particle physics, astronomy, economics, and many more. Extensive studies have been conducted in order to understand the mechanism behind such intricate and yet geometrically simple behaviors. As shown in [136, 128, 43, 62, 36, 10, 98], these

emergent behaviors are steady-states of various types of collective dynamics, and they can be qualitatively studied when the governing equations are known beforehand. However, if only the short-time trajectories of the dynamics are observed, it may be challenging to make accurate predictions about the emergent behaviors of the observed dynamics without prior knowledge of the governing equations. We offer a learning approach to overcome this difficulty by first discovering the governing equations from the observational data, and then use the estimated equations for large-time prediction. Research on discovering governing equations of dynamical systems has enjoyed a long history in the science and engineering community; it can be traced back to the earlier work of Lagrange, Laplace and Gauss [126]. Among the many inspiring studies, the lengthy discovery of gravity had immense impact. In 1605, Kepler announced his first law of planetary motion, from his work on showing Mars' elliptical orbit based on Tycho Brahe's observational data. Based on Kepler's first law and the assumption that gravity has a parametric form, namely $\frac{1}{r^p}$, Newton formulated his law of universal gravitation, i.e., that gravity has the form $1/r^2$, in 1687. Our learning approach can re-discover the $1/r^2$ form of the law of universal gravitation in a highly efficient and precise manner without the assumption of gravitation having a parametric form and planetary motion being elliptical, for details see Sec. 3.7.

In the third chapter, we consider the generalization of this model to evolve on Riemannian manifolds, this generalization allows for richer models and we are able to maintain the optimality of our estimators under mild assumptions on the system and the manifold. It is a fundamental challenge to learn the governing equations of interacting agent systems. Often, agents are either associated with state variables which belong to non-Euclidean spaces, e.g., phase variables considered in various Kuramoto models [78, 127], or constrained to move on non-Euclidean spaces, for example [3]. This has motivated a growing body of research considering interacting agent systems on various manifolds [81, 26, 117], including opinion dynamics [6],

flocking models [3] and a classical aggregation model [60]. Further recent approaches for interacting agents on manifolds include [144, 124]. In this work, we offer a nonparametric and inverse-problem-based learning approach to infer the governing structure of interacting agent dynamics, in the form of $\dot{\boldsymbol{X}} = \boldsymbol{f}(\boldsymbol{X})$, constrained on Riemannian manifolds. Our method is different from others introduced to learn ODEs and PDEs from observations, that aim at inferring $\boldsymbol{f}$, and would be cursed by the high-dimension of the state space of $\boldsymbol{X}$. Instead, we exploit the form of the function $\boldsymbol{f}$, special to interacting agent systems, which is determined by an underlying interaction kernel function ϕ of one variable only, and learn ϕ, with minimal assumptions on ϕ. By exploiting invariance of the equations under permutation of the agents as well as the radial symmetry of ϕ, we are able to overcome the curse of dimensionality, while most other approaches (Bayesian, spars regression, neural networks) are cursed by the dimension of the state space. We also demonstrate how our approach can perform transfer learning in section 4.5.

In the final chapter, we give a detailed exploration of real data that looks to determine whether our model can learn an effective form of gravity based purely on the trajectories of planets and the Sun in our Solar System. Precisely modeling and predicting celestial motion has had impacts on human societies across the world and initiated studies that led to fundamental developments in Physics [88, 113]. These discoveries of fundamental Physics usually worked hand in hand with novel mathematical tools to provide explanations of the observation data. In 1687, Newton presented the famous $\frac{1}{r^2}$-form of gravity; but, in 1845, Le Verrier found that the perihelion precession rate of Mercury could not be explained by Newton's theory of gravity. It took another 70 years, and the development of Riemannian geometry and relativity, for Einstein to explain that the discrepancy was due to the effect of the curvature of spacetime around the Sun. Einstein's theory has since been applied to celestial motion well beyond the Solar system. See [133, 8, 105, 2] and references

therein. With the rapid development of advanced observation technologies, statistics and machine learning have enabled us to analyze big data sets and discover novel patterns that are nearly impossible for a human to identify. It is well suited to play a complementary role to traditional physical reasoning in the pursuit of discovering fundamental Physics [72, 27]. There has been extensive research applying machine learning in science (especially in Physics) and engineering, examples include: learning PDEs [9, 118], governing equations [30], behavior in Biology [33], and fluid mechanics [110, 69]. Further examples include: many-body problems in quantum systems [28], mean field games [116], meteorology [68], dynamical systems [39, 24, 142, 18], and discrete field theory [106]. By combining effective theory [140] and Machine Learning, our data-driven modeling can provide interpretable and meaningful physical models which can be used to model observations with very high accuracy, preserving not only the geometric properties of the trajectories, but also localized dynamical features such as perihelion precession rates. These results help to confirm the learning theory and demonstrate the effectiveness of our learning algorithm on real data coming from second order interacting agent systems of fundamental physical interest.

Chapter 2

Theoretical Results

2.1 Introduction

Our learning approach discovers the governing laws of a particular subset of dynamical systems of the form,

$$\dot{\boldsymbol{Y}}(t) = \boldsymbol{F}_{\boldsymbol{\phi}^{EA},\boldsymbol{\phi}^{\xi}}(\boldsymbol{Y}(t)), \quad \boldsymbol{Y}(0) = \boldsymbol{Y}_0 \in \mathbb{R}^D, \quad t \in [0,T].$$

The learning problem is to infer the right hand side function $\boldsymbol{F}_{\boldsymbol{\phi}^{EA},\boldsymbol{\phi}^{\xi}}$ from observations $\{\boldsymbol{Y}_{t_l}^{(m)}, \dot{\boldsymbol{Y}}_{t_l}^{(m)}\}_{m=1,l=1}^{M,L}$ of the dynamical system, where m indexes different trajectories, started from an initial conditions (IC) sampled i.i.d. from a measure $\boldsymbol{\mu}^{\boldsymbol{Y}}$ on the state space. Here M is the total number of trajectories observed, with each trajectory forming a single observation (M plays a fundamental role in the learning theory where we study convergence as M varies); L refers to the number of observations at different times along each trajectory. Throughout this work, m will index the trajectories $1,\ldots,M$ and l will index the points in time $1,\ldots,L$. The main difficulties in establishing an effective theory of learning $\boldsymbol{F}_{\boldsymbol{\phi}^{EA},\boldsymbol{\phi}^{\xi}}$ are the *curse of dimensionality* caused by the dimension of $\boldsymbol{Y}$, which is $D = N(2d+1)$, where N is the number of agents, d the dimension of physical space; and the *dependence* of the observation data,

for example $\boldsymbol{Y}(t_{l+1})$ is a deterministic function of $\boldsymbol{Y}(t_l)$.

We present a learning approach based on exploiting the structure of collective dynamical systems and nonparametric estimation techniques (see [44, 132, 64, 55, 15]). A simplified form of our model equations, generalizing the first order models (see discussion in Appendix 2.12), is derived from Newton's second law and given by: for $i = 1, \ldots, N$

$$m_i \ddot{\boldsymbol{x}}_i(t) = \mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i) + \frac{1}{N}\sum_{i'=1}^{N} \phi^E(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|)(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)) + \phi^A(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|)(\dot{\boldsymbol{x}}_{i'}(t) - \dot{\boldsymbol{x}}_i(t)). \tag{2.1.1}$$

Here, m_i is the mass of the i^{th} agent, $\boldsymbol{x}_i$ is its position, $\mathbf{F}^{\dot{\boldsymbol{x}}}$ is a non-collective force, and $\phi^E, \phi^A : \mathbb{R}^+ \to \mathbb{R}$ are known as the *interaction kernels*.

To use the trajectory data to derive estimators, we consider appropriate hypothesis spaces in which to build our estimators, measures adapted to the dynamics, norms, and other performance metrics, and ultimately an inverse problem built from these tools. More specifically, let $\widehat{\boldsymbol{\phi}}^{EA}$ denote the direct sum of the kernels $\widehat{\boldsymbol{\phi}}^E \oplus \widehat{\boldsymbol{\phi}}^A$ (for the notation, see section 2.3), and define our estimator as

$$\widehat{\boldsymbol{\phi}}^{EA} := \underset{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}}{\arg\min}\ \boldsymbol{\mathcal{E}}_M^{EA}(\boldsymbol{\varphi}^{EA}),$$

where $\boldsymbol{\mathcal{E}}_M^{EA}$ is an empirical error functional depending on the observation data, $\boldsymbol{\mathcal{H}}^{EA}$ is a hypothesis space to search for our estimators, and based on the form of the error functional the estimator is calculated as the solution of a constrained least squares problem. Once we have obtained this estimated interaction kernel, we want to study its properties as a function of the amount of trajectory data we receive, which is the M trajectories sampled from different initial conditions from the same underlying system, each consisting of L time observations along the trajectory. Here we study

properties of the error functional, establish the uniqueness of its minimizers, and use the probability measures to define a dynamics-adapted norm to measure the error of our estimators over the hypothesis spaces. In comparing the estimators to the true interaction kernels, we first establish concentration estimates over the hypothesis space.

Our first main result is the strong asymptotic consistency of our learned estimators, as the number M of trajectories increases, which for the model (2.1.1) yields:

$$\lim_{M\to\infty} \|\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} = 0 \quad \text{with probability one,} \tag{2.1.2}$$

where $\boldsymbol{\rho}_T^{EA,L}$ is a dynamics-adapted measure on pairwise distances, and we use a weighted $\boldsymbol{L}^2$ space (see section 2.4, particularly (2.4.3)); see section 2.3 for the required definitions and section 2.4.3 for the full theorem. In fact, we also prove a stronger result that provides the rate of convergence. We achieve the minimax rate of convergence for any number of variables $\mathcal{V}$ in the interaction kernels. See section 2.4.4 for the full theorem, (see section 2.3 for relevant definitions) which is given by:

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\|\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}\right] \leq C\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+1}}. \tag{2.1.3}$$

In the case of model (2.1.1), $\mathcal{V} = 1$, as in the results for first-order systems [19, 89, 90].

This means that our estimators converge at the same rate in M as the best possible estimator (up to a logarithmic factor) one could construct when the initial conditions are randomly sampled from some underlying initial condition distribution denoted $\boldsymbol{\mu}^{\boldsymbol{Y}}$ throughout this work, see (section 2.4.3).

To solve the inverse problem, we give a detailed discussion of an essential link between these three aspects, the notion of coercivity of the system - detailed in section 2.4.2. Coercivity plays a key role in the approximation properties, the algorithm design, and the learning theory. We also present numerical examples, see also the

detailed numerical study in [146], which help to explain why the particular norms we define are the right choice, as well as show excellent performance on complex dynamical systems, in section 2.5.

The chapter is structured as follows. The first part of the chapter describes the model, learning framework, inference problem, and the basic tools needed for the learning theory. These ideas are all explained in detail in sections 2.2-2.4. If one wishes to quickly jump to the theoretical sections, and then refer back to the definitions as needed, we have provided tables 2.1, 2.3 which explains the model equations and outlines the definitions and concepts needed for the learning theory and general theoretical results, respectively. The theoretical part of the chapter (sections 2.4.2-2.4.5) discusses fundamental questions of identifiability and solvability of the inverse problem, consistency, and rate of convergence of the estimators, and the ability to control trajectory error of the evolved trajectories using our estimators. Some key highlights of our theoretical contributions are described in 2.3.4, with full details in the corresponding sections. Lastly, we consider applications in section 2.5, as well as have many additional proofs and details in appendices 2.7-2.14.

2.1.1 Comparison with existing work

Our general method is a non-parametric, inverse-problem-based approach to infer the interaction kernels from observations of trajectory data, especially within short-time periods. In [19], a convergence study of learning unknown interaction kernels from observation of first-order models of homogeneous agents was done for increasing N, the number of agents. The estimation problem with N fixed, but the number of trajectories M varying, for first-order and second-order models of heterogeneous agents was numerically studied in [89] and learning theory on these first-order models was developed in [90, 84]. Further extensions of the model and algorithm to more complicated second-order systems, with particular emphasis on emergent collective

behaviors, was discussed in [146]. A big data application to real celestial motion ephemerides is developed and discussed in [94]. In this work, we provide a rigorous learning theory covering the models presented in [146], as well as the second-order models introduced in [89]. We consider generalizations of the models in [146], to include models with higher-dimensional interaction kernels, that do not depend only on pairwise distances. Compared to the theories studied in [90, 84], our theory focuses on second-order models with interaction kernels of the form $\phi^E(r)r + \phi^A(r)\dot{r}$ (with r and $\dot{r}$ representing norms of differences of positions and, respectively, velocities of pairs of agents); additionally, we discuss the identifiability and separability of ϕ^E and ϕ^A from the sum.

More generally, applying machine learning to the sciences has experienced tremendous growth in recent years, a small selection of general applications related to the ideas in this work include: learning PDEs ([9, 118, 80]), modeling dynamical systems ([58, 12, 83]), governing equations ([30, 143]), biology ([33]), fluid mechanics ([110, 69]), many-body problems in quantum systems ([28]), mean-field games ([116]), meteorology ([68]), and dynamical systems ([39, 24, 142, 18]). These, and the references therein, give a flavor of the diverse range of applications. A vast literature exists in the context of learning dynamical systems. In the case of a general nonlinear dynamical system, symbolic regression has been developed to learn the underlying form of the equations from data, see [17, 119]. Sparse regression techniques which use an extremely large collection of functions, often containing most major mathematical functions, are fit to the data with a sparsity condition that only allows a few terms to appear in the final model. Detailed study and development of these approaches can be found for SINDy in ([23, 115, 22]), a LASSO-type penalty ([70, 73]), and sparse Bayesian regression ([145]). Other approaches consider multiscale methods, statistical mechanics, or force-based models, see [7, 14]. Deep learning has also been applied to learn dynamical systems, for ODEs see [109, 114] and for PDEs see [107, 108, 87], as well as the references

therein.

The majority of the earliest work in inferring interaction kernels in systems of the type (2.1.1), (2.2.2) occurred in the Physics literature, going back to the works of Newton. From the viewpoint of purely data-driven analysis of the equations, requiring limited or no physical reasoning, foundational work on estimating interaction laws includes [91, 74]. In these works, the interaction kernels are assumed to be in the span of a known family of functions and parameters are estimated. In statistics, the problem of parameter estimation in dynamical systems from observations is classical, e.g. [135, 21, 85, 25, 111]. The question of identifiability of the parameter emerges, see e.g. [52, 95]. Our work is closely related to this viewpoint but our parameter is now infinite-dimensional, with identifiability discusses in section 2.4.2.

There are many techniques which can be used to tackle the high-dimensionality of the data set: sparsity assumptions, dimension reduction, reduced-order modeling, and machine learning techniques trained using gradient-based optimization. The dependent nature of the data prevents traditional regression-based approaches, see the discussion in [90], but many of the approaches above successfully address this. Our work, however, exploits the interacting-agent structure of collective dynamical systems, which is driven by a collection of two-body interactions where each interaction depends only on pairwise data between the states of agents, as in (2.1.1). With this structure in mind, we are able to reduce the ambient dimension of the data $N(2d+1)$ to the dimension of the variables in the interaction kernels, which is independent of N. We also naturally incorporate the dependence in the data in an appropriate manner by considering trajectories generated from different initial conditions.

Our theoretical results focus on the joint learning of $\boldsymbol{\phi}^E, \boldsymbol{\phi}^A$ that takes into account their natural weighted direct sum structure that is described in the following sections, which is different from the learning theory on single $\boldsymbol{\phi}^E$'s considered in [89, 90]. The current theoretical framework is not able to conclusively show that $\boldsymbol{\phi}^E$ and $\boldsymbol{\phi}^A$ can be

learned separately; however we demonstrate in various numerical experiments that by learning $\boldsymbol{\phi}^E$ and $\boldsymbol{\phi}^A$ jointly, we still achieve strong performance. Finally, we note that the first-order theory developed in [90] is a special case of our second-order theory, see details in appendix 2.12.

2.2 Model description

In order to motivate the choice of second-order models considered in this chapter, we begin our discussion with a simple second-order model derived from classical mechanics. Let us consider a closed system of N homogeneous agents (or particles) equipped with a certain type of Lagrangian $L(t)$ in the form

$$L(t) = \frac{1}{2}\sum_{i=1}^{N} m_i \left\|\dot{\boldsymbol{x}}_i(t)\right\|^2 - \frac{1}{2N}\sum_{i,i'=1}^{N} U(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|), \quad i = 1, \dots, N\,.$$

Here U is a potential energy depending on pairwise distance. From the Lagrange equation, $\frac{d}{dt}\partial_{\dot{\boldsymbol{x}}_i} L = \partial_{\boldsymbol{x}_i} L$, we obtain the second-order collective dynamics model

$$m_i \ddot{\boldsymbol{x}}_i(t) = \frac{1}{N}\sum_{i'=1}^{N} \phi^E(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|)(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)), \quad i = 1, \dots, N\,. \tag{2.2.1}$$

Here, $\phi^E(r) = \frac{U'(r)}{r}$ represents an energy-based interaction between agents. We are assuming a regularity condition on ϕ^E, i.e. $\phi^E(0)0 = 0$. For example, the choice $U(r) = \frac{NGm_{i'}m_i}{r}$ corresponds to Newton's gravity model.

In order to incorporate a wider spectrum of behaviors, we add alignment-based interactions, which enable the alignment of velocities (so that short-range repulsion, mid-range alignment, and long arrange attraction are all present), auxiliary state variables describing internal states of agents (emotion, excitation, phases, etc.), and non-collective forces (interaction with the environment). We also allow for heterogeneous systems, consisting of agents belonging to K disjoint types $\{C_k\}_{k=1}^K$, with N_k

being the number of agents of type k, grouped in the index subset C_k. In summary, the systems we consider have the form

$$\begin{cases} m_i\ddot{\boldsymbol{x}}_i(t) &= \mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i(t), \dot{\boldsymbol{x}}_i(t), \xi_i(t)) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \Big[\phi^E_{k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^E_{ii'}(t))(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)) \\ & \qquad\qquad + \phi^A_{k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^A_{ii'}(t))(\dot{\boldsymbol{x}}_{i'}(t) - \dot{\boldsymbol{x}}_i(t))\Big] \\ \dot{\xi}_i(t) &= \mathbf{F}^{\xi}(\boldsymbol{x}_i(t), \dot{\boldsymbol{x}}_i(t), \xi_i(t)) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^{\xi}_{k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^{\xi}_{ii'}(t))(\xi_{i'}(t) - \xi_i(t)) \end{cases} \tag{2.2.2}$$

for $i = 1, \ldots, N$, where $k_i \in \{1, \cdots, K\}$ is the index of the agent type of the agent i. The interaction kernels $\phi^E_{kk'}, \phi^A_{kk'}, \phi^{\xi}_{kk'}$ are in general different for interacting agents of different types, and they not only depend on the pairwise distance $r_{ii'}(t) = \|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|$, but also on other pairwise features, $\boldsymbol{s}^E_{ii'}(t), \boldsymbol{s}^A_{ii'}(t), \boldsymbol{s}^{\xi}_{ii'}(t)$. Note the implicit dependence of t for these feature variables. For example, the interactions between birds may depend on the field of vision, not just the distance between pairs of birds. We will often suppress the explicit dependence on time t when it is clear from the context. The unknowns, for which we will construct estimators, in these equations, are the functions $\phi^E_{k_i k_{i'}}$, $\phi^A_{k_i k_{i'}}$ and $\phi^{\xi}_{k_i k_{i'}}$; everything else is assumed given.

Table 2.1 gives a detailed explanation for the definition of the variables used in (2.2.2). We note that in what follows, the notation $\{E, A, \xi\}$ attached to a map/function/etc... means that there is one of those maps/functions/etc. for each element in the set $\{E, A, \xi\}$. It is a convenient way to avoid excessive repetition of similar definitions.

Variable	Definition
i, i'	index of agent, from $1, \dots, N$
m_i	mass of agent i
$\boldsymbol{x}_i(t), \dot{\boldsymbol{x}}_i(t), \ddot{\boldsymbol{x}}_i(t) \in \mathbb{R}^d$	position/velocity/acceleration vector of agent i at time t
$\xi_i, \dot{\xi}_i$	auxiliary variable, and its derivative
$\|\cdot\|$	Euclidean norm in $\mathbb{R}^d$
K	number of agent types
k, k'	index for agent types, ranging in $\{1, \dots, K\}$
N_k	number of agents in type k
k_i	element of $\{1, \dots, K\}$ indicating the type of agent i
C_k	subset of $\{1, \dots, N\}$ consisting of indices of the agents of type k
$\phi_{kk'}$	influence of any agent of type k onto any agent of type k'
$\mathbf{F}^{\dot{\boldsymbol{x}}}, \mathbf{F}^{\xi}$	non-collective forces affecting $\ddot{\boldsymbol{x}}_i$ and $\dot{\xi}_i$, respectively
ϕ^E, ϕ^A, ϕ^ξ	energy, alignment, and environment-based interaction kernels respectively
$\mathcal{F}$	Common Feature Map (CFM), $\mathcal{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}, \xi, \boldsymbol{x}', \dot{\boldsymbol{x}}', \xi') : \mathbb{R}^{4d+2} \to \mathbb{R}^p$
$\pi_{kk'}^{\{E,A,\xi\}}$	Projection map, give the feature combination from CFM
$\boldsymbol{s}_{(k,k')}^{\{E,A,\xi\}}$	Feature map, $\pi_{kk'}^{\{E,A,\xi\}} \circ \mathcal{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}, \xi, \boldsymbol{x}', \dot{\boldsymbol{x}}', \xi') : \mathbb{R}^{4d+2} \to \mathbb{R}^{p_{(k,k')}^{\{E,A,\xi\}}}$
$\boldsymbol{s}_{ii'}^{\{E,A,\xi\}}(t)$	Feature evaluation, $\boldsymbol{s}_{(k_i,k_{i'})}^{\{E,A,\xi\}}(\boldsymbol{x}_i(t), \dot{\boldsymbol{x}}_i(t), \xi_i(t), \boldsymbol{x}_{i'}(t), \dot{\boldsymbol{x}}_{i'}(t), \xi_{i'}(t)) \in \mathbb{R}^{p_{k_i k_{i'}}^{\{E,A,\xi\}}}$

Table 2.1: Notation for the variables in (2.2.2).

The specific instances of the feature map $\mathcal{F}$ together with corresponding projections $\pi_{kk'}^{\{E,A,\xi\}}$ include a variety of systems that have found a wide range of applications in physics, biology, ecology, and social science; see the examples in the chart below. We assume that the function $\mathcal{F}$ is Lipschitz and known, and so are all the $\pi_{kk'}^{\{E,A,\xi\}}$'s. The Lipschitz assumption is sufficient to ensure the well-posedness of the system and will also be used to control the trajectory error, and of course implies that the feature maps $\boldsymbol{s}_{(k,k')}^{\{E,A,\xi\}}$ are all Lipschitz. The function $\mathcal{F}$ is a uniform way to collect all of the different variables (functions of the inputs) used across any of the (k, k') pairs over all of the E, A, ξ functions in the system. This uniformity is helpful when discussing the rate of convergence, among other places. Examples of where this generality matters emerge naturally, say when one has a different number of variables across interaction kernels for different pairs (k, k'), or when the energy and alignment kernels depend on r and then additional but distinct other variables. From this uniform set of variables, we then project to arrive at the relevant function $\boldsymbol{s}_{(k,k')}^{\{E,A,\xi\}}$ for each pair (and each of the elements of the wildcard). Lastly, we can then evaluate this map at the specific pair

	Properties											
Model	ϕ^E	ϕ^A	m_i	$\mathbf{F}^{\dot{\boldsymbol{x}}}$	ϕ^ξ	$\mathbf{F}^\xi$	$K>1$	$\boldsymbol{s}^E$	$\boldsymbol{s}^A$	$\boldsymbol{s}^\xi$	$\mathcal{V}$	$\mathcal{V}^\xi$
Anticipation Dynamics	■	■	■					■	■		2	
Celestial Mechanics	■		■				■				1	
Cucker-Smale		■	■								1	
Fish Milling 2D	■		■	■							1	
Fish Milling 3D	■		■	■							1	
Flocking w. Ext. Poten.	■	■	■	■							1	
Phototaxis		■	■	■	■	■					1	1
Predator-Swarm (2^{nd} Order)	■		■	■			■				1	
Lennard-Jones	■					■					1	
Opinion Dynamics	■										1	
Predator-Swarm (1^{st} Order)	■						■				1	
Synchronized Oscillator		■		■	■	■		■		■	2	2

Table 2.2: Summary of the models studied in this work and in [89, 90, 146, 94]

of agents (i, i'), that leads to the feature evaluation, $\boldsymbol{s}_{ii'}^{\{E,A,\xi\}}$ which is the expression used in the model equation (2.2.2).

The model class (2.2.2) is quite large. We will consider several different concrete example in section 2.5.2. We summarize how those examples, and others, map to the model class in table 2.2, with a shaded (respectively: empty) cell indicating that the model has (respectively: has not) that characteristic. A numeric value indicates this is the number of unique variables, $\mathcal{V}, \mathcal{V}^\xi$ used within the EA or ξ portions of the system. The number of these unique variables specifies the dimension in the minimax convergence rate, see section 2.4.4.

Our second-order model equations cover the first-order models considered in [89, 90, 146] as special cases (see Appendix 2.12), but they are a significantly larger class: even when written as a first-order system in more variables, they are a strict generalization of the previous first-order models. Furthermore, the dynamical characteristics produced by second-order models are much richer and can model more complicated collective motions and emergent behavior of the agents.

2.3 Inference problem and learning approach

In this section, we first introduce the problem of inferring the interaction kernels from observations of trajectory data and give a brief review and generalization of the learning approach proposed in the works [89] and [146].

2.3.1 Preliminaries and notation

We vectorize the model in (2.2.2) in order to obtain a more compact description. We let $\boldsymbol{v}_i(t) := \dot{\boldsymbol{x}}_i(t)$ and

$$\boldsymbol{X}_t := \begin{bmatrix} \boldsymbol{x}_1(t) \\ \vdots \\ \boldsymbol{x}_N(t) \end{bmatrix} \in \mathbb{R}^{Nd}, \quad \boldsymbol{V}_t := \begin{bmatrix} \boldsymbol{v}_1(t) \\ \vdots \\ \boldsymbol{v}_N(t) \end{bmatrix} \in \mathbb{R}^{Nd}, \quad \boldsymbol{\Xi}_t := \begin{bmatrix} \xi_1(t) \\ \vdots \\ \xi_N(t) \end{bmatrix} \in \mathbb{R}^{N}.$$

We introduce the weighted norm

$$\|\boldsymbol{Z}\|_{\mathcal{S}}^2 := \sum_{i=1}^{N} \frac{1}{N_{k_i}} \|\boldsymbol{z}_i\|^2 \tag{2.3.1}$$

for $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^T, & \dots, & \boldsymbol{z}_N^T \end{bmatrix}^T$ with each $\boldsymbol{z}_i \in \mathbb{R}^d$ or $\mathbb{R}$. Here $\|\cdot\|$ is the same norm used in the construction of pairwise distance data for the interaction kernels (typically, the Euclidean norm). The weight factor $1/N_{k_i}$ is introduced so that different types of agents of different types are overall weighted equally, which is important in the estimation phase, especially in the case when the number of agents of different types is highly non-uniform. The model (2.2.2) becomes

$$\begin{cases} \vec{m} \circ \ddot{\boldsymbol{X}}_t &= \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) + \mathbf{f}^{\phi^E}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) + \mathbf{f}^{\phi^A}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) \\ \dot{\boldsymbol{\Xi}}_t &= \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) + \mathbf{f}^{\phi^\xi}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t)\,. \end{cases}$$

Here $\vec{m} = \begin{bmatrix} m_1, & \dots, & m_N \end{bmatrix}^T \in \mathbb{R}^N$, $\circ$ is the Hadamard product, and we use boldface fonts to denote the vectorized form of our estimators (with some once-for-all-fixed ordering of the pairs $(k,k')_{k,k'=1,\dots,K}$):

$$\boldsymbol{\phi}^E = [\phi^E_{kk'}]^K_{k,k'=1}, \quad \boldsymbol{\phi}^A = [\phi^A_{kk'}]^K_{k,k'=1}, \quad \boldsymbol{\phi}^\xi = [\phi^\xi_{kk'}]^K_{k,k'=1}\,, \tag{2.3.2}$$

and of the non-collective force:

$$\mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\Xi}) := \left[\mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i,\dot{\boldsymbol{x}}_i,\xi_i)\right]^T_{i=1,\dots,N},$$
$$\mathbf{f}^{\boldsymbol{\phi}^E} := \left[\sum_{i'=1}^N \frac{1}{N_{\mathfrak{k}_{i'}}}\phi^E_{\mathfrak{k}_i\mathfrak{k}_{i'}}(r_{ii'},\boldsymbol{s}^E_{ii'})(\boldsymbol{x}_{i'}-\boldsymbol{x}_i)\right]^T_{i=1,\dots,N}$$

both of which are vectors in $\mathbb{R}^{Nd}$. We omit the analogous definitions for $\mathbf{f}^{\boldsymbol{\phi}^A}$ and $\mathbf{f}^{\boldsymbol{\phi}^\xi}$. We also use the shorthand:

$$\boldsymbol{\phi}^{EA} := \boldsymbol{\phi}^E \oplus \boldsymbol{\phi}^A\,, \tag{2.3.3}$$

to denote the element of the direct sum of the function spaces containing $\boldsymbol{\phi}^E, \boldsymbol{\phi}^A$.

2.3.2 Problem setting

Our observation data is given by $\{\mathbf{Y}^{(m)}_{t_l}, \dot{\mathbf{Y}}^{(m)}_{t_l}\}^{M,L}_{m=1,l=1}$ for $0 = t_1 < t_2 < \cdots < t_L = T$. Here $\dot{\boldsymbol{Y}}_t = [\boldsymbol{y}_1^T(t), \cdots, \boldsymbol{y}_N^T(t)]$ and $\boldsymbol{y}_i(t) = \left[\boldsymbol{x}_i^T(t), \dot{\boldsymbol{x}}_i^T(t), \xi_i(t)\right]^T$, and m indexes the M different trajectories, each generated by the system (2.1) with the unknown set of interaction kernels, i.e. $\boldsymbol{\phi}^E, \boldsymbol{\phi}^A, \boldsymbol{\phi}^\xi$, with initial conditions $\{\boldsymbol{Y}^{(m)}(0)\}_{m=1,\dots,M}$ drawn i.i.d from $\boldsymbol{\mu}^{\boldsymbol{Y}}$, a probability measure defined on the space $\mathbb{R}^{N(2d+2)}$. We use a superscript (m) to denote that the variable is calculated from the data from that m^{th} trajectory. The objective is to construct estimators $\widehat{\boldsymbol{\phi}}^E, \widehat{\boldsymbol{\phi}}^A, \widehat{\boldsymbol{\phi}}^\xi$ the unknown interaction kernels given these observations.

2.3.3 Loss functionals

For simplicity, we only consider equidistant observation points: $t_l - t_{l-1} = h$ for $l = 2, \dots, L$; the proposed estimator is easily extended to the case non-equispaced time points. Following and extending [89, 90, 146], we consider the empirical error functional (recall the shorthand notation (2.3.3))

$$\begin{aligned}\boldsymbol{\mathcal{E}}_M^{EA}(\boldsymbol{\varphi}^{EA}) &:= \frac{1}{LM}\sum_{l=1,m=1}^{L,M}\Big\|\ddot{\boldsymbol{X}}_{t_l}^{(m)} - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_{t_l}^{(m)}, \boldsymbol{V}_{t_l}^{(m)}, \boldsymbol{\Xi}_{t_l}^{(m)}) \\ &\qquad - \mathbf{f}^{\boldsymbol{\varphi}^E}(\boldsymbol{X}_{t_l}^{(m)}, \boldsymbol{V}_{t_l}^{(m)}, \boldsymbol{\Xi}_{t_l}^{(m)}) - \mathbf{f}^{\boldsymbol{\varphi}^A}(\boldsymbol{X}_{t_l}^{(m)}, \boldsymbol{V}_{t_l}^{(m)}, \boldsymbol{\Xi}_{t_l}^{(m)})\Big\|_{\mathcal{S}}^2, \\ \boldsymbol{\mathcal{E}}_M^{\xi}(\boldsymbol{\varphi}^{\xi}) &:= \frac{1}{LM}\sum_{l=1,m=1}^{L,M}\Big\|\dot{\boldsymbol{\Xi}}_{t_l} - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_{t_l}^{(m)}, \boldsymbol{V}_{t_l}^{(m)}, \boldsymbol{\Xi}_{t_l}^{(m)}) - \mathbf{f}^{\boldsymbol{\varphi}^{\xi}}(\boldsymbol{X}_{t_l}^{(m)}, \boldsymbol{V}_{t_l}^{(m)}, \boldsymbol{\Xi}_{t_l}^{(m)})\Big\|_{\mathcal{S}}^2.\end{aligned} \tag{2.3.4}$$

The estimators of interaction kernels are defined as the minimizers of the error functionals $\boldsymbol{\mathcal{E}}_M^{EA}$ and $\boldsymbol{\mathcal{E}}_M^{\xi}$ over suitably chosen finite-dimensional function spaces $\boldsymbol{\mathcal{H}}^{EA}$ and $\boldsymbol{\mathcal{H}}^{\xi}$:

$$\widehat{\boldsymbol{\phi}}^{EA} = \operatorname*{arg\,min}_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}} \boldsymbol{\mathcal{E}}_M^{EA}(\boldsymbol{\varphi}^{EA}), \qquad \widehat{\boldsymbol{\phi}}^{\xi} = \operatorname*{arg\,min}_{\boldsymbol{\varphi}^{\xi}\in\boldsymbol{\mathcal{H}}^{\xi}} \boldsymbol{\mathcal{E}}_M^{\xi}(\boldsymbol{\varphi}^{\xi}). \tag{2.3.5}$$

2.3.4 Overview of theoretical contributions

We focus on the regime where L is fixed but $M \to \infty$. We provide a learning theory that answers the fundamental questions:

- **Quantitative description of estimator errors**. We will introduce measures to describe how close the estimators are to the true interaction kernels, that lead to novel dynamics-adapted norms. See section 2.4.
- **Identifiability of kernels**. We will establish the existence and uniqueness of the estimators as well as relate the solvability of our inverse problem to a fundamental coercivity property. See section 2.4.2.
- **Consistency and optimal convergence rate of the estimators**. We will

prove theorems on strong consistency and optimal minimax rates of convergence of the estimators, which exploit the separability of the learning on the energy and alignment from the learning on the environment variable. See section 2.4.3.

- **Trajectory Prediction** We prove a theorem that describes the performance of the estimated dynamics using the estimated kernels compared to the true dynamics. Our result demonstrates how the expected supremum error (over the entire time interval) of our trajectories is controlled by the norm of the difference between the true and estimated kernels, further justifying our choice of norms and estimation procedure. See section 2.4.5.

The papers [146, 89, 90] have applied this learning approach to a variety of systems and the extensive numerical simulations demonstrate the effectiveness of the approach. However, theoretical guarantees of the proposed approach for second order systems had not been developed and will be the main focus of this chapter. Our theory includes the first-order theory in [90] as a special case, as discussed in Appendix 2.12.

2.3.5 Function spaces

We begin by describing some basic ideas about measures and function spaces. Consider a compact or precompact set $\mathcal{U} \subset \mathbb{R}^p$ for some p; the infinity norm is defined as $\|h\|_\infty := \operatorname{ess\,sup}_{x\in\mathcal{U}} |h(x)|$, and $L^\infty(\mathcal{U})$ as the space of real valued functions defined on $\mathcal{U}$ with finite ∞-norm. A key function space we need to consider is, $C_c^{k,\alpha}(\mathcal{U})$, for $k \in \mathbb{N}$, $0 < \alpha \leq 1$, defined as the space of compactly supported, k-times continuously differentiable functions with a k-th derivative that is Hölder continuous of order α. We can then consider vectorizations of these spaces over agent types as

$$\boldsymbol{L}^\infty(\mathcal{U}) := \bigoplus_{k,k'=1,1}^{K,K} L^\infty(\mathcal{U})\,, \text{ endowed with the norm } \|\boldsymbol{f}\|_\infty := \max_{k,k'} \|f_{kk'}\|_\infty\,, \forall \boldsymbol{f} \in \boldsymbol{L}^\infty(\mathcal{U})\,.$$

Similarly, we consider direct sums of measures, with corresponding vectorized function spaces, in particular L^2 (see section 2.4.1).

We now define a suitable function class for the interaction kernels in the model (2.2.2). A simple model is that the agents get farther and farther apart, they eventually should have no influence on each other. For each pair (k, k'), $k, k' = 1, \dots K$ we define the admissible space

$$\mathcal{K}_{kk'}^{\{E,A,\xi\}} := L^\infty([R_{kk'}^{\min}, R_{kk'}^{\max}] \times \mathbb{S}_{kk'}^{\{E,A,\xi\}}) \quad , \quad \boldsymbol{\mathcal{K}}^{\{E,A,\xi\}} := \bigoplus_{k,k'=1,1}^{K,K} \mathcal{K}_{kk'}^{\{E,A,\xi\}} , \tag{2.3.6}$$

where we remind the reader that the $\{E, A, \xi\}$ notation means, in this case, that there is an admissible space for each element of the set $\{E, A, \xi\}$. Here, $R_{kk'}^{\min}, R_{kk'}^{\max}$ are the minimum or maximum, respectively, possible interaction radius for agents in $C_{k'}$ influencing agents in C_k. Similarly, $\mathbb{S}_{kk'}^E, \mathbb{S}_{kk'}^A, \mathbb{S}_{kk'}^\xi$ are compact sets in $\mathbb{R}^{p_{kk'}^E}, \mathbb{R}^{p_{kk'}^A}, \mathbb{R}^{p_{kk'}^\xi}$ which contain the ranges of the feature maps, $\boldsymbol{s}_{kk'}^E, \boldsymbol{s}_{kk'}^A$ and $\boldsymbol{s}_{kk'}^\xi$. We will also need the sets:

$$\mathbf{S}^{\{E,A,\xi\}} := \prod_{k,k'} \mathbb{S}_{kk'}^{\{E,A,\xi\}} \quad , \quad \mathbf{R} := \prod_{k,k'} [R_{kk'}^{\min}, R_{kk'}^{\max}] \quad , \quad R := \max_{k,k'} R_{kk'}^{\max}. \tag{2.3.7}$$

Notice that all interaction kernels are supported on the interval of pairwise distance $[0, R]$.

We denote the distribution of the initial conditions by $\boldsymbol{\mu}^{\boldsymbol{Y}}$. This measure is unknown and is the source of randomness in our system. It is a product measure of three measures $\boldsymbol{\mu}^{\boldsymbol{X}}$, $\boldsymbol{\mu}^{\boldsymbol{V}}$, $\boldsymbol{\mu}^{\xi}$, all also unknown, that represent the distribution on the initial positions, velocities, and environment variables, respectively. Specifically, we define,

$$\boldsymbol{\mu}^{\boldsymbol{Y}} := \begin{bmatrix} \mu^{\boldsymbol{X}} \\ \mu^{\boldsymbol{V}} \\ \mu^{\boldsymbol{\Xi}} \end{bmatrix} \tag{2.3.8}$$

It reflects that we will observe trajectories which start at different initial conditions, but that evolve from the same dynamical system. For example, in our numerical experiments we will choose $\boldsymbol{\mu}^{\boldsymbol{Y}}$ to be uniform over a system-dependent compact set.

We let

$$R_{\dot{x}} := \sup_{\boldsymbol{Y}(0)\sim\boldsymbol{\mu}^{\boldsymbol{Y}}} \sup_{t\in[0,T]} \max_{i,i'} \|\dot{\boldsymbol{x}}_i(t) - \dot{\boldsymbol{x}}_{i'}(t)\|\,, \tag{2.3.9}$$

$$R_{\xi} := \sup_{\boldsymbol{Y}(0)\sim\boldsymbol{\mu}^{\boldsymbol{Y}}} \sup_{t\in[0,T]} \max_{i,i'} \|\xi_i(t) - \xi_{i'}(t)\|\,, \tag{2.3.10}$$

and we assume that both of these quantities are finite. A sufficient condition is that the measures $\boldsymbol{\mu}^{\boldsymbol{V}}, \boldsymbol{\mu}^{\xi}$ (specifying the distribution of the initial conditions on the velocities and the environment variable are compactly supported, which follows by the assumptions on the interaction kernels below and that we only consider finite final time T.

First, we define the following vectorized function spaces, which we call *admissible sets*,

$$\boldsymbol{\mathcal{K}}^{\{E,A,\xi\}}_{S_{\{E,A,\xi\}}} := \Big\{ \left(\phi^{\{E,A,\xi\}}_{kk'}\right)^{K,K}_{k,k'=1,1} : \forall k,k' = 1,\dots,K\,, \phi^{\{E,A,\xi\}}_{kk'} \in C^{0,1}\left([R^{\min}_{kk'}, R^{\max}_{kk'}] \times \mathbb{S}^{\{E,A,\xi\}}_{kk'}\right)\,, \tag{2.3.11}$$

$$\left\|\phi^{\{E,A,\xi\}}_{kk'}\right\|_{\infty} + \mathrm{Lip}\left[\phi^{\{E,A,\xi\}}_{kk'}\right] \le S_{\{E,A,\xi\}} \Big\}\,.$$

Next, we will assume that the interaction kernels live in the corresponding admissible sets, namely,

$$\boldsymbol{\phi}^E \in \boldsymbol{\mathcal{K}}^E_{S_E}, \qquad \boldsymbol{\phi}^A \in \boldsymbol{\mathcal{K}}^A_{S_A}, \qquad \boldsymbol{\phi}^{\xi} \in \boldsymbol{\mathcal{K}}^{\xi}_{S_{\xi}}. \tag{2.3.12}$$

The admissibility assumptions (2.3.12) allow us to establish properties such as existence and uniqueness of solutions to (2.2.2) as well as to have control on the trajectory errors in finite time $[0,T]$. It further allows us to show regularity and absolute continuity with respect to Lebesgue measure of the appropriate performance

measures defined in section 2.4.1.

When estimating the EA part of the system, we will consider the direct sum admissible space, for $S_{EA} \geq \max\{S_E, S_A\}$,

$$\boldsymbol{\mathcal{K}}_{S_{EA}}^{EA} := \boldsymbol{\mathcal{K}}_{S_E}^{E} \oplus \boldsymbol{\mathcal{K}}_{S_A}^{A} \tag{2.3.13}$$

In the learning approach, we will consider hypothesis spaces that we will search in order to estimate the various interaction kernels. The hypothesis spaces corresponding to $\{\phi_{kk'}^{\{E,A,\xi\}}\}$ are denoted as $\{\mathcal{H}_{kk'}^{\{E,A,\xi\}}\}$ and we vectorize them as,

$$\boldsymbol{\mathcal{H}}^{\{E,A,\xi\}} := \bigoplus_{k,k'=1,1}^{K,K} \mathcal{H}_{kk'}^{\{E,A,\xi\}}. \tag{2.3.14}$$

Analogous to our simplified notation for $\boldsymbol{\phi}^{EA}, \boldsymbol{\varphi}^{EA}$ described in (2.3.3), we define the direct sum of the hypothesis spaces as,

$$\boldsymbol{\mathcal{H}}^{EA} := \boldsymbol{\mathcal{H}}^{E} \oplus \boldsymbol{\mathcal{H}}^{A} \tag{2.3.15}$$

We will consider specific choices for the hypothesis spaces during the learning theory and numerical algorithm sections.

2.3.6 Algorithm for constructing the interaction kernel estimators

Let $\mathcal{H}_{kk'}^{\boldsymbol{x}}$ be a finite dimensional function spaceof dimension $n_{kk'}^{E}$ with basis functions given by piecewise polynomials whose degree will be chosen later (other type of basis functions are also possible, e.g., clamped B-splines as shown in [89]). It is built on uniform partitions of $[R_{kk'}^{\min,\text{obs}}, R_{kk'}^{\max,\text{obs}}]$ where $R_{kk'}^{\min,\text{obs}}/R_{kk'}^{\max,\text{obs}}$ is the minimum/maximum interacting radius for agents in type k' influencing agents in type k,

derived from the observation data. Similar construction is done for $\mathcal{H}^{\boldsymbol{x}}$ with dimension $n^A_{kk'}$. We write the candidate $\varphi^E_{kk'}, \varphi^A_{kk'}$ as a linear combination of the basis functions:

$$\varphi^E_{kk'}(r, \boldsymbol{s}^E) = \sum_{\eta^E_{kk'}=1}^{n^E_{kk'}} \alpha^E_{k,k',\eta^E_{kk'}} \psi^{\boldsymbol{x}}_{k,k',\eta^E_{kk'}}(r, \boldsymbol{s}^E),$$
$$\varphi^A_{kk'}(r, \dot{r}, \boldsymbol{s}^A) = \sum_{\eta^A_{kk'}=1}^{n^A_{kk'}} \alpha^A_{k,k',\eta^A_{kk'}} \psi^{\boldsymbol{x}}_{k,k',\eta^A_{kk'}}(r, \dot{r}, \boldsymbol{s}^A), .$$

Substituting this linear combination back into (2.3.4), we obtain a system of linear equations,

$$A^{EA}_M \vec{\alpha}^{EA} = \vec{b}^{EA}_M ,$$

and minimizing the empirical loss functional corresponds to solving this system in the last square sense. Here, $\vec{\alpha}^{EA} \in \mathbb{R}^{n^{EA}} = \begin{bmatrix} (\vec{\alpha}^E)^T & (\vec{\alpha}^A)^T \end{bmatrix}^T$ with $\vec{\alpha}^E$ and $\vec{\alpha}^A$ being the collection of $\alpha^E_{k,k',\eta^E_{kk'}}$ or $\alpha^A_{k,k',\eta^A_{kk'}}$ respectively. Moreover, $A^{EA}_M \in \mathbb{R}^{n^{EA} \times n^{EA}}$ and $\vec{b}^{EA}_M \in \mathbb{R}^{n^{EA}}$. See Sec. 2.14 for full details.

The total computational complexity is detailed as follows: MLN^2 for computing pairwise data, $MLd(n^{EA})^2$ for constructing the learning matrix and right hand side vector, and $(n^{EA})^3$ for solving the linear system, hence the total computing time is $MLN^2 + MLd(n^{EA})^2 + (n^{EA})^3$. Assuming that a tensor-grid construction of $\mathcal{H}^{\boldsymbol{x}}_{kk'}$ is used, and the optimal n_* with

$$n_* = \mathcal{O}\left(\left(\frac{M}{\log M}\right)^{\frac{1}{2s+\mathcal{V}}}\right) \approx M^{\frac{1}{2s+\mathcal{V}}}.$$

is used in each dimension, we have $n^{EA} = 2n_*^{\mathcal{V}} \approx 2M^{\frac{\mathcal{V}}{2s+\mathcal{V}}}$; then we obtain the total computing time in terms of M as follows

$$\text{Comp. Time} = MLN^2 + 4LdM^{\frac{2\mathcal{V}}{2s+\mathcal{V}}+1} + 8M^{\frac{3\mathcal{V}}{2s+\mathcal{V}}}$$

In the special case of $s = 1$ (Lipscthiz functions) and $\mathcal{V} = 1$, we have

$$\text{Comp. Time} = MLN^2 + 4LdM^{\frac{2}{3}+1} + 8M \approx M^{\frac{2}{3}+1}.$$

It is slightly super-linear in M.

Similar computational complexity analysis on solving $A^{\xi}\vec{\alpha}^{\xi} = \vec{b}^{\xi}$ also shows that the computational cost is slightly super linear in M when $s = 1$ and $\mathcal{V} = 1$.

The overall memory storage needed for the learning problem is $MLN(d(5 + n^{EA} + n^{\xi}) + 3)$, with $MLN(4d + 2)$ for storing the trajectory data, $MLNd(n^{EA} + n^{\xi})$ (here $n^{EA} = n^E + n^A$) for learning matrices, and $MLN(d + 1)$ for right hand side vectors. Hence if $M \gg \mathcal{O}(1)$, we can consider parallelization in m in order to reduce the overhead memory, ending up with $M_{\text{per core}}\big(LN(d(5 + n^{EA} + n^{\xi}) + 3)\big)$ with $M_{\text{per core}} = \frac{M}{n_{\text{cores}}}$. The final storage of A and $\vec{b}$ only needs $n^{EA}(n^{EA} + 1) + n^{\xi}(n^{\xi} + 1)$.

2.4 Learning theory

2.4.1 Probability measures and weighted L^2 for measuring learning performance

The interaction kernels depend on $(r, \dot{r}, \boldsymbol{s}^E, \boldsymbol{s}^A, \boldsymbol{s}^{\xi})$, and to measure distances between estimated interaction kernels and true interaction kernels, we consider a natural set of probability measures and corresponding weighted L^2 spaces. These generalize the

constructions of [19, 89, 90]. For each interacting pair (k, k'), we let

$$\begin{aligned}
\rho_T^{EA,k,k'}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) &:= \mathbb{E}_{\boldsymbol{Y}_0 \sim \boldsymbol{\mu}^{\boldsymbol{Y}}} \frac{1}{TN_{kk'}} \int_{t=0}^{T} \sum_{\substack{i \in C_k, i' \in C_{k'} \\ i \neq i'}} \delta_{ii',t}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A)\, dt \\
\rho_T^{EA,L,k,k'}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) &:= \mathbb{E}_{\boldsymbol{Y}_0 \sim \boldsymbol{\mu}^{\boldsymbol{Y}}} \frac{1}{LN_{kk'}} \sum_{l=1}^{L} \sum_{\substack{i \in C_k, i' \in C_{k'} \\ i \neq i'}} \delta_{ii',t_l}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) \\
\rho_T^{EA,L,M,k,k'}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) &:= \frac{1}{MLN_{kk'}} \sum_{l,m=1}^{L,M} \sum_{\substack{i \in C_k, i' \in C_{k'} \\ i \neq i'}} \delta_{ii',t_l,m}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A)
\end{aligned} \tag{2.4.1}$$

where $N_{kk'} = N_k N_{k'}$ for $k \neq k'$ and $N_{kk'} = \binom{N_k}{2}$ for $k = k'$, and we used the following shorthand notation for the Dirac measures:

$$\begin{aligned}
\delta_{ii',t}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) &:= \delta_{r_{ii'}(t), \boldsymbol{s}^E_{ii'}(t), \dot{r}_{ii'}(t), \boldsymbol{s}^A_{ii'}(t)}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) \\
\delta_{ii',t,m}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A) &:= \delta_{r^{(m)}_{ii'}(t), \boldsymbol{s}^{E,(m)}_{ii'}(t), \dot{r}^{(m)}_{ii'}(t), \boldsymbol{s}^{A,(m)}_{ii'}(t)}(r, \boldsymbol{s}^E, \dot{r}, \boldsymbol{s}^A)\,.
\end{aligned}$$

The measure $\rho_T^{EA,L,k,k'}$ is the discrete counterpart of $\rho_T^{EA,k,k'}$ with the continuous average over $[0, T]$ replaced by the average over the observation times $0 = t_1 < \ldots < t_L = T$. $\rho_T^{EA,L,M,k,k'}$ can be computed from observations and converges to $\rho_T^{EA,L,k,k'}$ as $M \to \infty$.

We also consider the marginal distributions

$$\rho_T^{E,k,k'}(r, \boldsymbol{s}^E) := \int_{\dot{r}} \int_{\boldsymbol{s}^A} \rho_T^{EA,k,k'}\, d\boldsymbol{s}^A\, d\dot{r} \quad , \quad \rho_T^{A,k,k'}(r, \dot{r}, \boldsymbol{s}^A) := \int_{\boldsymbol{s}^E} \rho_T^{EA,k,k'}\, d\boldsymbol{s}^E \tag{2.4.2}$$

and $\rho_T^{E,L,k,k'}(r, \boldsymbol{s}^E)$, $\rho_T^{E,L,M,k,k'}(r, \boldsymbol{s}^E)$, $\rho_T^{A,L,k,k'}(r, \dot{r}, \boldsymbol{s}^A)$, $\rho_T^{A,L,M,k,k'}(r, \dot{r}, \boldsymbol{s}^A)$ defined analogously as above. The empirical measures, $\rho_T^{E,L,M,k,k'}, \rho_T^{A,L,M,k,k'}$, are the ones used in the actual algorithm to quantify the learning performances of the estimators $\hat{\phi}^E_{kk'}$ and $\hat{\phi}^A_{kk'}$ respectively. They are also crucial in discussing the separability of $\hat{\phi}^E_{kk'}$ and $\hat{\phi}^A_{kk'}$.

For ease of notation, we introduce the following measures to handle the heterogeneity of the system, and which are used to describe error over all of the pairs (k, k').

$$\boldsymbol{\rho}_T^{EA,L} = \bigoplus_{k,k'=1,1}^{K,K} \rho_T^{EA,L,kk'}, \quad \boldsymbol{\rho}_T^{EA} = \bigoplus_{k,k'=1,1}^{K,K} \rho_T^{EA,kk'}, \quad \boldsymbol{L}^2\left(\boldsymbol{\rho}_T^{EA,L}\right) = \bigoplus_{k,k'=1,1}^{K,K} L^2\left(\rho_T^{EA,L,kk'}\right) \tag{2.4.3}$$

Similar definitions apply for measures related to learning the ξ-based interaction kernels, see Appendix 2.13. We discuss some key properties of the measures in Appendix 2.10.

We now discuss the performance measures for the estimated interaction kernels. We use weighted L^2-norms (with mild abuse of notation, we omit the weight from the notation) based on the dynamics-adapted measures introduced above (with analogous definitions for the measures corresponding to finite L):

$$\begin{aligned}
\left\|\hat{\phi}_{kk'}^E - \phi_{kk'}^E\right\|_{L^2(\rho_T^{E,k,k'})}^2 &:= \int_{(r,\boldsymbol{s}^E)} (\hat{\phi}_{kk'}^E(r,\boldsymbol{s}^E) - \phi_{kk'}^E(r,\boldsymbol{s}^E))^2 r^2 \, d\rho_T^{E,k,k'} \\
\left\|\hat{\phi}_{kk'}^A - \phi_{kk'}^A\right\|_{L^2(\rho_T^{A,k,k'})}^2 &:= \int_{(r,\dot{r},\boldsymbol{s}^A)} (\hat{\phi}_{kk'}^E(r,\dot{r},\boldsymbol{s}^A) - \phi_{kk'}^E(r,\dot{r},\boldsymbol{s}^A))^2 \dot{r}^2 \, d\rho_T^{A,k,k'} \\
\left\|\hat{\phi}_{kk'}^{EA} - \phi_{kk'}^{EA}\right\|_{L^2(\rho_T^{EA,k,k'})}^2 &:= \int_{r,\boldsymbol{s}^E,\dot{r},\boldsymbol{s}^A} \Big[(\hat{\phi}_{kk'}^E(r,\boldsymbol{s}^E) - \phi_{kk'}^E(r,\boldsymbol{s}^E))r \\
&\qquad + (\hat{\phi}_{kk'}^A(r,\dot{r},\boldsymbol{s}^A) - \phi_{kk'}^A(r,\dot{r},\boldsymbol{s}^A))\dot{r}\Big]^2 \, d\rho_T^{EA,k,k'} .
\end{aligned} \tag{2.4.4}$$

Our learning theory focuses on minimizing the difference between $\hat{\phi}_{kk'}^E \oplus \hat{\phi}_{kk'}^A$ and $\phi_{kk'}^E \oplus \phi_{kk'}^A$ in the joint norm given by (2.4.4). As long as the joint norm is small, our estimators produce faithful approximations of the right hand side function of the original system and trajectories. However, it does not necessarily imply that both $\hat{\phi}_{kk'}^E - \phi_{kk'}^E$'s and $\hat{\phi}_{kk'}^A - \phi_{kk'}^A$'s are small in their corresponding energy- and alignment-based norms, since the joint norm is a weaker norm. It would be interesting to study if there is any equivalence between these two norms, but the problem appears to be quite delicate. The theoretical investigation is still ongoing.

Now, we have all the tools needed to establish a theoretical framework: dynamics

induced probability measures, performance measurements in appropriate norms, and loss functionals. These will allow us to discuss the convergence properties of our estimators. Full details of the numerical algorithm are given in Appendix 2.14.

Notational summary

A summary of the learning theory notation introduced in sections 2.3.1, 2.3, and the notation above, is given below in table 2.3.

Notation	Definition	Ref
M	number of trajectories	Sec. 2.1
L	number of times in $[0,T]$ for each trajectory	Sec. 2.2
$\boldsymbol{Y}(t)$	full state space vector containing $\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t$	Sec. 2.2
$\{E,A,\xi\}$	wildcard, means the notation applies for all 3 variables	Sec. 2.2
$\|\boldsymbol{X}\|_{\mathcal{S}}$	$\sum_{i=1}^{N} \frac{1}{N_{k_i}} \|\boldsymbol{x}_i\|^2$	(2.3.1)
$\boldsymbol{\mu}^{\boldsymbol{Y}}$	distribution on the initial conditions $\boldsymbol{Y}(0)$	Sec. 2.3.5
$\boldsymbol{\phi}^{\{E,A,\xi\}} = (\phi_{kk'}^{\{E,A,\xi\}})_{k,k'}$	vectorized true E,A,ξ interaction kernels	(2.3.2)
$\boldsymbol{\varphi}^{\{E,A,\xi\}} = (\varphi_{kk'}^{\{E,A,\xi\}})_{k,k'}$	$\boldsymbol{\varphi}^{\{E,A,\xi\}} \in \boldsymbol{\mathcal{H}}^{\{E,A,\xi\}}$ with $\varphi_{kk'}^{\{E,A,\xi\}} \in \mathcal{H}_{kk'}^{\{E,A,\xi\}}$	(2.3.2)
EA	shorthand denoting energy and alignment part of system	(2.3.3)
$\boldsymbol{\phi}^{EA}$	represents the joint function $\boldsymbol{\phi}^E \oplus \boldsymbol{\phi}^A \in \boldsymbol{\mathcal{H}}^{EA}$	(2.3.3)
$\|\boldsymbol{Y}\|_{\mathcal{Y}}^2$	$\|\boldsymbol{X}\|_{\mathcal{S}}^2 + \|\boldsymbol{V}\|_{\mathcal{S}}^2 + \|\boldsymbol{\Xi}\|_{\mathcal{S}}^2$.	(2.4.24)
$\mathbf{S}^{\{E,A,\xi\}}$	$\prod_{k,k'} \mathbb{S}_{kk'}^{\{E,A,\xi\}}$	(2.3.7)
$\mathbf{R}$	$\prod_{k,k'} [R_{kk'}^{\min}, R_{kk'}^{\max}]$	(2.3.7)
R	$\max_{k,k'} R_{kk'}^{\max}$	(2.3.7)
$\boldsymbol{\mathcal{K}}_{S_{\{E,A,\xi\}}}^{\{E,A,\xi\}}$	admissible spaces for the E,A,ξ kernels	(2.3.11)
$\mathcal{H}_{kk'}^{\{E,A,\xi\}}$	the hypothesis spaces for $\phi_{kk'}^{\{E,A,\xi\}}$	(2.3.14)
$\boldsymbol{\mathcal{H}}^{\{E,A,\xi\}} = \oplus_{kk'} \mathcal{H}_{kk'}^{\{E,A,\xi\}}$	the hypothesis spaces for $\boldsymbol{\phi}^{\{E,A,\xi\}}$	(2.3.14)
$\boldsymbol{\mathcal{H}}^{EA}$	direct sum of hypothesis spaces $\boldsymbol{\mathcal{H}}^E \oplus \boldsymbol{\mathcal{H}}^A$	(2.3.15)
$\boldsymbol{\mathcal{E}}_M^{EA}(\cdot), \boldsymbol{\mathcal{E}}_M^{\xi}(\cdot)$	empirical EA error functional, ξ error functional	Sec. 2.3.3
$\widehat{\boldsymbol{\phi}}_M^{EA} := \widehat{\boldsymbol{\phi}}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}^{EA}$	$\operatorname{argmin}_{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}} \boldsymbol{\mathcal{E}}_M^{EA}(\boldsymbol{\varphi}^{EA})$	(2.3.5)
$\widehat{\boldsymbol{\phi}}_M^{\xi} := \widehat{\boldsymbol{\phi}}_{L,M,\boldsymbol{\mathcal{H}}^{\xi}}^{\xi}$	$\operatorname{argmin}_{\boldsymbol{\varphi}^{\xi} \in \boldsymbol{\mathcal{H}}^{\xi}} \boldsymbol{\mathcal{E}}_M^{\xi}(\boldsymbol{\varphi}^{\xi})$	(2.3.5)
$\{\psi_{kk',p}^{\{E,A,\xi\}}\}_{p=1}^{n_{kk'}^{\{E,A,\xi\}}}$	basis for $\mathcal{H}_{kk'}^{\{E,A,\xi\}}$	Sec. 2.3.6
$\boldsymbol{\rho}_T^{EA}$, $\boldsymbol{\rho}_T^{\xi}$	measure for EA, ξ with continuous time, infinite trajectories	(2.4.3), 2.13
$\boldsymbol{\rho}_T^{EA,L}$, $\boldsymbol{\rho}_T^{\xi,L}$	measure for EA, ξ discrete in time, infinite trajectories	(2.4.3), 2.13
$\boldsymbol{L}^2\left(\boldsymbol{\rho}_T^{EA,L}\right)$	$\bigoplus_{k,k'=1,1}^{K,K} L^2\left(\rho_T^{EA,L,kk'}\right)$	(2.4.3)
$c_{\boldsymbol{\mathcal{H}}^{EA}}$, $c_{\boldsymbol{\mathcal{H}}^{\xi}}$	coercivity constant on the $\boldsymbol{\mathcal{H}}^{EA}$, $\boldsymbol{\mathcal{H}}^{\xi}$ hypothesis spaces	Def.2.4.1
$\boldsymbol{\mathcal{H}}_M^{EA}$,$\boldsymbol{\mathcal{H}}_M^{\xi}$	hypothesis spaces on EA,ξ depending on M	Sec. 2.4.3
$\mathcal{V}, \mathcal{V}^{\xi}$	Dimension for minimax convergence rates	(2.4.16)
A_M^{EA}, A_M^{ξ}	Learning matrices for the inverse problem	Sec. 2.5
$\mathcal{N}(\boldsymbol{\mathcal{H}}, \delta)$	δ-covering number, under the ∞-norm, of a set $\boldsymbol{\mathcal{H}}$	[134]

Table 2.3: Notation used throughout the chapter

2.4.2 Identifiability of kernels from data

In this section we introduce a technical condition, called coercivity condition, on the dynamical system that relates to the well-posedness (solvability and uniqueness of the solution) of the inverse problem and plays a key role in the learning theory. We establish theorems in two directions:

1. showing how identifiability of the interaction kernels can be derived from the coercivity condition by relating the coercivity constant to the singular values of the learning matrices associated to our inverse problem, for both finitely and infinitely many trajectories;

2. establishing the coercivity condition for a wide class of dynamical systems of the form (2.2.2), under assumptions on the distribution $\boldsymbol{\mu}^{\boldsymbol{Y}}$ of the initial conditions. Our numerical experiments suggest that the coercivity condition holds even more generally.

For the remainder of the chapter, we will make the following assumptions on the hypothesis spaces used in the learning approach:

Assumption 2.4.1. *$\boldsymbol{\mathcal{H}}^{EA}$ is a compact convex subset of $\boldsymbol{\mathcal{K}}^{EA}_{S_{EA}} := \boldsymbol{\mathcal{K}}^{E}_{S_E} \oplus \boldsymbol{\mathcal{K}}^{A}_{S_A}$ (see 2.3.13)*

This implies that the infinity norm of all elements in $\boldsymbol{\mathcal{H}}^{EA}$ is bounded above by $\max\{S_E, S_A\}$, and we assume that a constant $S_{EA} \geq \max\{S_E, S_A\}$ is known.

Assumption 2.4.2. *$\boldsymbol{\mathcal{H}}^{\xi}$ is a compact convex subset of $\boldsymbol{\mathcal{K}}^{\xi}_{S_\xi}$ (see 2.3.11).*

This implies the elements of $\boldsymbol{\mathcal{H}}^{\xi}$ have $\infty-$norm bounded above by S_ξ, and we assume that a constant $S_0 \geq S_\xi$ is known. It is easy to see that $\boldsymbol{\mathcal{H}}^{EA}$ can be naturally embedded as a compact subset of $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})$ and that $\boldsymbol{\mathcal{H}}^{\xi}$ can be naturally embedded as a compact subset of $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi,L})$ (recall these measures are defined in section 2.4.1).

Assumptions 2.4.1, 2.4.2 ensure the existence of minimizers to the loss functionals $\boldsymbol{\mathcal{E}}_M^{EA}, \boldsymbol{\mathcal{E}}_M^{\xi}$ defined in (2.3.4) and (2.3.4), which will be proven in Appendix 2.8.

In order to ensure learnability we introduce a coercivity condition, that generalizes that introduced in [19] and studied in [89, 90, 84]. In fact, for the second-order systems considered here, we will have two coercivity conditions, one for the energy and alignment terms and the other for the ξ variable. These conditions ensure that the minimizers to the error functionals are unique, and second that when the expected error functional is small, then the distance from the estimator to the true kernels is small in the appropriate $\boldsymbol{\rho}_T$ norm. Due to their connection to the error functional and the learnability of the interaction kernels, coercivity plays an important role in the theorems of Section 2.4.3.

Definition 2.4.1 (Coercivity condition)**.** *For the dynamical system* (2.2.2) *observed at time instants* $0 = t_1 < t_2 < \cdots < t_L = T$ *and with initial condition distributed* $\boldsymbol{\mu}^{\boldsymbol{Y}}$ *on* $\mathbb{R}^{(2d+1)N}$*, it satisfies the coercivity condition on the hypothesis space* $\boldsymbol{\mathcal{H}}^{EA}$ *with constant* $c_{\boldsymbol{\mathcal{H}}^{EA}}$ *if*

$$c_{\boldsymbol{\mathcal{H}}^{EA}} := \inf_{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA} \setminus \{\mathbf{0}\}} \frac{\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\left\| \mathbf{f}_{\boldsymbol{\varphi}^{EA}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) \right\|_{\mathcal{S}}^2 \right]}{\|\boldsymbol{\varphi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}} > 0. \tag{2.4.5}$$

Similarly, the system satisfies the coercivity condition on the hypothesis space $\boldsymbol{\mathcal{H}}^{\xi}$ *with constant* $c_{\boldsymbol{\mathcal{H}}^{\xi}}$ *if*

$$c_{\boldsymbol{\mathcal{H}}^{\xi}} := \inf_{\boldsymbol{\varphi}^{\xi} \in \boldsymbol{\mathcal{H}}^{\xi} \setminus \{\mathbf{0}\}} \frac{\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\left\| \mathbf{f}_{\boldsymbol{\varphi}^{\xi}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) \right\|_{\mathcal{S}}^2 \right]}{\|\boldsymbol{\varphi}^{\xi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi,L})}} > 0. \tag{2.4.6}$$

Analogous definitions holds for continuous observations over the time interval $[0, T]$*, by replacing the average over observations at discrete times with an integral average over* $[0, T]$*.*

In the following, we prove the coercivity condition on general compact sets of $\boldsymbol{L}^2([0,R],\boldsymbol{\rho}_T^{EA,L})$ under suitable hypotheses. Our result is independent of N, which implies that the finite sample bounds of Theorem 2.4.7 can be dimension free – in that the coercivity constant has no dependence on N. This result implies that coercivity may be a fundamental property of the dynamical system, including in the mean field regime ($N \to \infty$).

Identifiability from coercivity

By choosing the hypothesis space to be compact and convex, we are able to show that the error functional has a unique minimizer. However, many possible bases exist that could potentially yield good performance (in terms of the error functional and $\boldsymbol{L}^2$ error to the true kernel). We want to choose a basis such that the regression matrix, A_M^{EA} defined in Appendix 2.14, is well-conditioned – and thus an estimator can be learned that will have good performance (in terms of the error functional and $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})$ error to the true kernel). In the proposition below we establish two results in this direction. The key for both results is that the basis is chosen to be orthonormal in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})$, versus the naive choice of basis in the underlying direct sum of $\boldsymbol{L}^\infty$ spaces that the interaction kernels live in. The first result is theoretical and shows that, under appropriate assumptions on the basis, the minimal singular value of the expected ($M \to \infty$) regression matrix (denoted A_∞^{EA}) equals the coercivity constant. The second result is critical for the practical implementation, as for each finite but large enough M it shows that the minimal singular value of the regression matrix A_M^{EA} is lower bounded in terms of the coercivity constant with high probability. These results demonstrate quantitatively that the regression matrix is well-conditioned if the coercivity constant of $\boldsymbol{\mathcal{H}}^{EA}$ well-separated from 0, and the importance of choosing the hypothesis space, and a basis thereof, to ensure this property, if at all possible for the dynamical system.

To ease the notation, we introduce the bilinear functional $\langle\langle \cdot, \cdot \rangle\rangle$ on $\boldsymbol{\mathcal{H}}^{EA} \times \boldsymbol{\mathcal{H}}^{EA}$, defined by

$$\begin{aligned}
&\langle\langle \boldsymbol{\varphi}_1^{EA}, \boldsymbol{\varphi}_2^{EA} \rangle\rangle := \\
&\frac{1}{L} \sum_{l,i=1}^{L,N} \frac{1}{N_{k_i}} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \Bigg[\Bigg\langle \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \Big(\varphi^E_{1,k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^E_{ii'}) \boldsymbol{r}_{ii'}(t) + \varphi^A_{1,k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^A_{ii'}) \dot{\boldsymbol{r}}_{ii'}(t) \Big), \\
&\qquad \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \Big(\varphi^E_{2,k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^E_{ii'}) \boldsymbol{r}_{ii'}(t) + \varphi^A_{2,k_i k_{i'}}(r_{ii'}(t), \boldsymbol{s}^A_{ii'}) \dot{\boldsymbol{r}}_{ii'}(t) \Big) \Bigg\rangle \Bigg]
\end{aligned} \tag{2.4.7}$$

for any $\boldsymbol{\varphi}_1^{EA} = (\varphi^E_{1,kk'} \oplus \varphi^A_{1,kk'})^{K,K}_{k,k'=1,1} \in \boldsymbol{\mathcal{H}}^{EA}$, and $\boldsymbol{\varphi}_2^{EA} = (\varphi^E_{2,kk'} \oplus \varphi^A_{2,kk'})^{K,K}_{k,k'=1,1} \in \boldsymbol{\mathcal{H}}^{EA}$. For every pair (k, k') let $(\psi^E_{kk',i} \oplus \psi^A_{kk',i})_{i=1}^{n_{kk'}}$ be a basis of

$$\mathcal{H}^{EA}_{kk'} \subset L^\infty([0, R] \times \mathbb{S}^E_{kk'}) \oplus L^\infty([0, R] \times \mathbb{S}^A_{kk'})$$

satisfying the orthonomality and boundedness conditions

$$\langle \psi^{EA}_{kk',p}, \psi^{EA}_{kk',p'} \rangle_{L^2(\rho_T^{EA,L,kk'})} = \delta_{p,p'}, \quad \|\psi^{EA}_{kk',p}\|_\infty \leq S_{EA}. \tag{2.4.8}$$

We note that multivariable basis functions arise naturally in this setting due to the model. For example, a tensor product basis of splines or piecewise polynomials can be used. The $n_{kk'}$ notation allows multivariable functions, different choices for the number of basis functions across pairs (k, k'), and a different number of basis functions within a pair with respect to the underlying coordinates of the tensor product.

By convention, we use the lexicographic ordering to order within pairs (k, k') (with order $r, \boldsymbol{s}^E, \boldsymbol{s}^A$), and then across pairs (with the lexicographic ordering on pairs of integers). Set $\mathbf{n} = \sum_{k,k'} n_{kk'} = dim(\boldsymbol{\mathcal{H}}^{EA})$; then for any function $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$, we can write

$$\boldsymbol{\varphi}^{EA} = \sum_{p=1}^{\mathbf{n}} a_p \boldsymbol{\psi}^{EA}_p.$$

Under the setting above, we have the following relationship between the coercivity constant and the minimal singular value of the empirical and expected learning matrix:

Proposition 2.4.3. *Consider the matrices*

$$A_\infty^{EA} = \left(\langle\langle \boldsymbol{\psi}_p^{EA}, \boldsymbol{\psi}_{p'}^{EA} \rangle\rangle\right)_{p,p'} \in \mathbb{R}^{\mathbf{n}\times\mathbf{n}}, \quad A_\infty^{\xi} = \left(\langle\langle \boldsymbol{\psi}_p^{\xi}, \boldsymbol{\psi}_{p'}^{\xi} \rangle\rangle\right)_{p,p'} \in \mathbb{R}^{\mathbf{n}_\xi\times\mathbf{n}_\xi},$$

and choose the hypothesis spaces as $\boldsymbol{\mathcal{H}}^{EA} = \text{span}\{\boldsymbol{\psi}_p^E \oplus \boldsymbol{\psi}_p^A\}_{p=1}^{\mathbf{n}}$ *and* $\boldsymbol{\mathcal{H}}^{\xi} = \text{span}\{\boldsymbol{\psi}_p^{\xi}\}_{p=1}^{\mathbf{n}_\xi}$ *as above. Then*

$$\sigma_{\min}(A_\infty^{EA}) = c_{\boldsymbol{\mathcal{H}}^{EA}} \quad , \quad \sigma_{\min}(A_\infty^{\xi}) = c_{\boldsymbol{\mathcal{H}}^{\xi}}, \tag{2.4.9}$$

with $c_{\boldsymbol{\mathcal{H}}^{EA}}, c_{\boldsymbol{\mathcal{H}}^{\xi}}$ *defined in* (2.4.5),(2.4.6). *Additionally, for large* M*, the smallest singular value of* A_M^{EA} *satisfies the inequality*

$$\sigma_{\min}(A_M^{EA}) \geq 0.8 c_{\boldsymbol{\mathcal{H}}^{EA}}$$

with probability at least $1 - 2\mathbf{n}\exp\left(-\frac{c_{\boldsymbol{\mathcal{H}}^{EA}}^2 M}{100\mathbf{n}^2 c_1^2 + \frac{20}{3}\cdot c_1 \cdot c_{\boldsymbol{\mathcal{H}}^{EA}}\cdot \mathbf{n}}\right)$ *with* $c_1 = 2K^4 \max\{R, R_{\dot{x}}\}^2 S_{EA}^2 + 2$. *Similarly, for large* M*, the smallest singular value of* A_M^{ξ} *satisfies the inequality*

$$\sigma_{\min}(A_M^{\xi}) \geq 0.8 c_{\boldsymbol{\mathcal{H}}^{\xi}}$$

with high probability at least $1 - 2\mathbf{n}_\xi \exp\left(-\frac{c_{\boldsymbol{\mathcal{H}}^{\xi}}^2 M}{100 n^2 c_1^2 + \frac{20}{3} c_2 \cdot c_{\boldsymbol{\mathcal{H}}^{\xi}} \cdot n}\right)$ *with* $c_2 = 2K^4 R_\xi^2 S_\xi^2 + 2$. *Therefore, with high probability, a system and its associated hypothesis space satisfying the coercivity condition, and for* M *sufficiently large, the inverse problem is uniquely solvable, with condition number controlled by the coercivity constant.*

Proof. We prove the result in the EA case, the proof of the results about the ξ part of the system being analogous. The orthonormality of the component functions given in (2.4.8), implies that $\langle \boldsymbol{\psi}_p^{EA}, \boldsymbol{\psi}_{p'}^{EA} \rangle_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} = \delta_{pp'}$. Expand $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$ in this basis

as $\boldsymbol{\varphi}^{EA} = \sum_{p=1}^{\mathbf{n}} a_p \boldsymbol{\psi}_p^{EA}$. Let the vector $v = (a_1, \ldots, a_{\mathbf{n}}) \in \mathbb{R}^{\mathbf{n}}$, and notice that

$$\begin{aligned}\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\left\|\mathbf{f}_{\boldsymbol{\varphi}^{EA}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l})\right\|_{\mathcal{S}}^2\right] &= \langle\!\langle \sum_{p=1}^{\mathbf{n}} a_p \boldsymbol{\psi}_p^{EA}, \sum_{p=1}^{\mathbf{n}} a_p \boldsymbol{\psi}_p^{EA} \rangle\!\rangle \\ &= v^T A_\infty^{EA} v \geq \sigma_{\min}(A_\infty^{EA})\|v\|^2 = \sigma_{\min}(A_\infty^{EA})\left\|\boldsymbol{\varphi}^{EA}\right\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}\end{aligned}$$

This lower bound is achieved by the singular vector corresponding to the singular value $\sigma_{\min}(A_\infty^{EA})$, so that by definition (2.4.5) we have that $\sigma_{\min}(A_\infty^{EA}) = c_{\boldsymbol{\mathcal{H}}^{EA}}$.

For the second statement, we consider the learning matrix A_M^{EA} (defined in section 2.14), from the observations from M trajectories. By construction, for each m, $A_\infty^{EA} = \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}[A^{EA,(m)}]$ and $\lim_{M\to\infty} A_M^{EA} = A_\infty^{EA}$ by the Strong Law of Large Numbers. Next we will derive some important properties of the learning matrix that will allow us to apply the matrix Bernstein inequality (see [131], Theorem 6.1.1, Corollary 6.1.2). Note that we will use the notation from this reference. First we note an elementary matrix analysis result (see [13] Problem III.6.13); for any two square matrices A, B, $\max_j |\sigma_j(A) - \sigma_j(B)| \leq \|A - B\|$. All norms in this proof are the spectral norm, unless otherwise specified. Thus if we get a concentration inequality of the form $\mathbb{P}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|A_\infty^{EA} - A_M^{EA}\| \geq t\}$ we will get the desired result relating the minimal singular values of A_M^{EA} to $c_{\boldsymbol{\mathcal{H}}^{EA}}$. First, notice that $\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}[A_M^{EA}] = A_\infty^{EA}$. Additionally, using the definition of the regression matrix, and our assumptions on the interaction kernels and the dynamics, we can bound every entry by $c_1 = 2K^4 \max\{R, R_{\dot{x}}\}^2 S_{EA}^2 + 2$. This immediately implies the bound

$$\|A_M^{EA} - A_\infty^{EA}\| \leq 2\mathbf{n}c_1.$$

Next, we upper bound the matrix variance statistic (in our case $Z = \sum_{k=1}^{M} S_k$ where

$S_k = \frac{1}{M} A^{EA,(k)}$), defined as

$$v(Z) = \max\{\|\mathbb{E}[(Z - \mathbb{E}Z)(Z - \mathbb{E}Z)^*\|, \|\mathbb{E}[(Z - \mathbb{E}Z)^*(Z - \mathbb{E}Z)\|\}.$$

Using a similar analysis to bound each entry of the matrices, we can arrive at the result that $v(Z) \leq 2\mathbf{n}^2 c_1^2$. Now, we apply the matrix Bernstein inequality to see that

$$\mathbb{P}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\Big\{\|A_M^{EA} - A_\infty^{EA}\| \geq t\Big\} \leq 2\mathbf{n} \exp\bigg(- \frac{c_{\boldsymbol{\mathcal{H}}^{EA}}^2 M}{100\mathbf{n}^2 c_1^2 + \frac{20 c_1 c_{\boldsymbol{\mathcal{H}}^{EA}}}{3}\mathbf{n}}\bigg).$$

Note that the M in the numerator comes because A_M^{EA} has a factor of $\frac{1}{M}$ on it. Lastly, choose $t = \frac{c_{\boldsymbol{\mathcal{H}}^{EA}}}{5}$, which together with the results above yield the desired inequality. □

From Proposition 2.4.3 we see that, for each hypothesis space $\mathcal{H}_{kk'}$, it is important to choose a basis that is well-conditioned in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L}), \boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi,L})$, instead of in the corresponding $\boldsymbol{L}^\infty$ spaces. If not, the learning matrices, defined in Appendix 2.14, A_M^{EA}, A_M^{ξ} may be ill-conditioned or even singular. This would lead to fundamental numerical challenges in solving for the (coefficients of the) interaction kernels. In order to mitigate these issues, one can use piecewise polynomials on a partition of the support of the empirical measure and/or use the pseudo-inverse with an adaptive tolerance for thresholding small singular values.

Discussions on the coercivity condition

The coercivity condition is key to the identifiability of the interaction kernels from data. It is determined by the distribution of the solution to the agent system and introduces constraints on the hypothesis space. For the second-order system, it is therefore related to the distribution $\boldsymbol{\mu}^{\boldsymbol{Y}}$ of the initial conditions, the true interaction kernels, and the non-collective force. The coercivity condition has been studied for first-order systems in [89, 90, 84]. For homogeneous systems, [89, 90] showed that the

coercivity condition holds true on any compact subset of the corresponding L^2 space for the case of $L = 1$. This result has been generalized to cover heterogeneous systems in [90] and a few examples of the stochastic homogeneous system including linear systems and nonlinear systems with stationary distributions for general L in [84].

In this chapter, we shall employ a similar idea as for first-order systems and extend the result to second-order systems. One key in the proof is to show the positiveness of integral operators that arise in the expectation in Eq. (2.4.5). We focus on a representative model of second-order homogeneous systems,

$$\begin{cases} m_i \ddot{\boldsymbol{x}}_i &= \mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N} (\phi^E(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\boldsymbol{x}_{i'} - \boldsymbol{x}_i) + \phi^A(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\dot{\boldsymbol{x}}_{i'} - \dot{\boldsymbol{x}}_i) \\ \dot{\xi}_i &= \mathbf{F}^{\xi}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N} \phi^{\xi}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\xi_{i'} - \xi_i) \end{cases} \tag{2.4.10}$$

which includes the first-order systems considered in [19, 89, 90] as special cases and various second-order system examples in [89, 146] as specific applications. We shall prove the coercivity condition holds true for the case $L = 1$:

Theorem 2.4.4. *Consider the system* (2.9.1) *at time* $t_1 = 0$ *with the initial distribution* $\mu_0^{\boldsymbol{Y}} = \begin{bmatrix} \mu_0^{\boldsymbol{X}} \\ \mu_0^{\dot{\boldsymbol{V}}} \\ \mu_0^{\boldsymbol{\Xi}} \end{bmatrix}$ *where* $\mu_0^{\boldsymbol{X}}$ *is exchangeable Gaussian with* $\operatorname{cov}(\boldsymbol{x}_i(t_1)) - \operatorname{cov}(\boldsymbol{x}_i(t_1), \boldsymbol{x}_j(t_1)) = \lambda I_d$ *for a constant* $\lambda > 0$, $\mu_0^{\dot{\boldsymbol{X}}}, \mu_0^{\boldsymbol{\Xi}}$ *are exchangeable with finite second moment, and they are independent of* $\mu_0^{\boldsymbol{X}}$. *Then*

$$\mathbb{E}_{\mu_0^{\boldsymbol{Y}}} \|\mathbf{f}_{\varphi^E \oplus \varphi^A}(\boldsymbol{X}_0, \boldsymbol{V}_0)\|_{\mathcal{S}}^2 \geq c_{1,N,\mathcal{H}^{EA}} \left\|\varphi^E \oplus \varphi^A\right\|_{L^2(\rho_T^{EA,1})},$$
$$\mathbb{E}_{\mu_0^{\boldsymbol{Y}}} \|\mathbf{f}_{\varphi^{\xi}}(\boldsymbol{X}_0, \boldsymbol{\Xi}_0)\|_{\mathcal{S}}^2 \geq c_{1,N,\mathcal{H}^{\xi}} \left\|\varphi^{\xi}\right\|_{L^2(\rho_T^{\xi,1})},$$

where

- $c_{1,N,\mathcal{H}^{EA}} \geq (\frac{N-1}{2N^2} + \frac{(N-1)(N-2)}{2N^2}c), c = \min\left\{c^E_{\mathcal{H}^{EA}}, c^A_{\mathcal{H}^{EA}} c_{\mu_0^{\dot{\boldsymbol{x}}}}\right\}$, *where* $c_{\boldsymbol{\mu}_0^{\dot{\boldsymbol{x}}}} = 1 - \frac{\mathbb{E}\langle \dot{\boldsymbol{x}}_i(0), \dot{\boldsymbol{x}}_{i'}(0)\rangle}{\mathbb{E}\|\dot{\boldsymbol{x}}_i(0)\|^2}$ *($i \neq i'$) and* $c^E_{\mathcal{H}^{EA}}$ *and* $c^A_{\mathcal{H}^{EA}}$ *are non-negative constants independent of N, and are strictly positive for compact* $\mathcal{H}^{EA}$ *of* $L^2(\rho_T^{EA,1})$.

- $c_{1,N,\mathcal{H}^{\xi}} \geq (\frac{N-1}{N^2} + \frac{(N-1)(N-2)}{N^2}c), c = c_{\mathcal{H}^{\xi}} c_{\mu_0^{\boldsymbol{\Xi}}}$ *with* $c_{\mu_0^{\boldsymbol{\Xi}}} = 1 - \frac{\mathbb{E}\langle \xi_i(0), \xi_{i'}(0)\rangle}{\mathbb{E}\|\xi_i(0)\|^2}$ *($i \neq i'$) and* $c_{\mathcal{H}^{\xi}}$ *is a non-negative constant independent of N, which is strictly positive for compact* $\mathcal{H}^{\xi}$ *of* $L^2(\rho_T^{\xi,1})$.

Note that in this case that the coercivity constant is independent of the number of agents N, and therefore not only will the convergence rate of our estimators be independent of the dimension $(2d+1)N$ of the phase space, but even the constants in front of the rate term are independent of N, see theorem 2.4.7. Our results extend those for first-order systems from [89, 90, 146]. The empirical numerical experiments on some second-order systems [146] support the idea that the coercivity condition is satisfied by large classes of second-order systems, and is "generally" satisfied for general L on suitable hypothesis spaces, with a constant independent of the number of agents N thanks to the exchangeability of the distribution of the initial conditions, and of the agents at any time t. The proof of the result above is given in Appendix 2.9.

2.4.3 Consistency and optimal convergence rate of estimators

The final preparatory results for our main theorems combine concentration with a union bound. Here we control the probability that the supremum of the difference between the expected and empirical normalized errors over the whole hypothesis space is large.

Concentration

Our first main result is a concentration estimate that relates the coercivity condition to an appropriate bias-variance tradeoff in our setting. Let $\mathcal{N}(\boldsymbol{\mathcal{H}}, \delta)$ be the δ-covering number, with respect to the ∞-norm, of the set $\boldsymbol{\mathcal{H}}$.

Theorem 2.4.5 (Concentration)**.** *Suppose that* $\boldsymbol{\phi}^{\{E,A,\xi\}} \in \boldsymbol{\mathcal{K}}^{\{E,A,\xi\}}_{S_{\{E,A,\xi\}}}$. *Consider a convex, compact (with respect to the* ∞*-norm) hypothesis spaces*

$$\boldsymbol{\mathcal{H}}^{EA}_M \subset \boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R}\times\mathbf{S}^E)\oplus\boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R}\times\mathbf{S}^A), \quad \boldsymbol{\mathcal{H}}^{\xi}_M \subset \boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R}\times\mathbf{S}^{\xi}),$$

bounded above by $S_0 \geq \max\{S_E, S_A, S_\xi\}$ *respectively. Additionally, assume that the coercivity conditions* (2.4.5),(2.4.6) *hold on* $\boldsymbol{\mathcal{H}}^{EA}_M$ *and* $\boldsymbol{\mathcal{H}}^{\xi}_M$*, respectively.*

Then for all $\epsilon > 0$*, with probability (with respect to* $\boldsymbol{\mu}^{\boldsymbol{Y}}$*) at least* $1-\delta$*, we have the estimates*

$$\begin{aligned} c_{\boldsymbol{\mathcal{H}}^{EA}_M}\|\widehat{\boldsymbol{\phi}}^{EA}_M - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{EA,L}_T)} &\leq 2\inf_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}_M}\|\boldsymbol{\varphi}^{EA}-\boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{EA,L}_T)} + 2\epsilon, \\ c_{\boldsymbol{\mathcal{H}}^{\xi}_M}\|\widehat{\boldsymbol{\phi}}^{\xi}_M - \boldsymbol{\phi}^{\xi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{\xi,L}_T)} &\leq 2\inf_{\boldsymbol{\varphi}^{\xi}\in\boldsymbol{\mathcal{H}}^{\xi}_M}\|\boldsymbol{\varphi}^{\xi}-\boldsymbol{\phi}^{\xi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{\xi,L}_T)} + 2\epsilon, \end{aligned} \tag{2.4.11}$$

provided that, for the first bound to hold,

$$M \geq \frac{1152 S_0^2 \max\{R, R_{\dot{x}}\}^2 K^4}{\epsilon c_{\boldsymbol{\mathcal{H}}^{EA}_M}}\left(\log\left(\mathcal{N}\left(\boldsymbol{\mathcal{H}}^{EA}_M, \frac{\epsilon}{48 S_0 \max\{R, R_{\dot{x}}\}^2 K^4}\right)\right) + \log\left(\frac{1}{\delta}\right)\right),$$

and similarly for the second inequality, using $\boldsymbol{\mathcal{H}}^{\xi}_M$.

Proof of Theorem 2.4.5. We start out by setting $\alpha = \frac{1}{6}$ in Proposition 2.8.11, which yields the tightest bound in the argument below. To ease the notation we let $\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}} = \widehat{\boldsymbol{\phi}}^{E}_{L,M,\boldsymbol{\mathcal{H}}^{EA}} \oplus \widehat{\boldsymbol{\phi}}^{A}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}$ and similarly for $\widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}$. From the Proposition, we have that

$$\sup_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}} \frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon} < \frac{1}{2}$$

holds true with probability

$$\mathcal{P} \geq 1 - \mathcal{N}\bigg(\boldsymbol{\mathcal{H}}^{EA}, \frac{\epsilon}{48 S_{EA} \max\{R, R_{\dot{x}}\}^2 K^4}\bigg) \exp\bigg(- \frac{c_{\boldsymbol{\mathcal{H}}^{EA}} M \epsilon}{1152 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^6}\bigg). \tag{2.4.12}$$

This immediately implies, by choosing $\boldsymbol{\varphi}^{EA} = \widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}$, that with probability $\mathcal{P}$

$$\mathcal{D}_\infty(\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}) < 2\mathcal{D}_M(\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}) + \epsilon\,.$$

By definition of $\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}$ as the minimizer of the empirical error functional $\boldsymbol{\mathcal{E}}^{EA}_M$, we see that

$$\mathcal{D}_M(\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}) = \boldsymbol{\mathcal{E}}^{EA}_M(\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}) - \boldsymbol{\mathcal{E}}^{EA}_M(\widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}) \leq 0,$$

and combining this result with equation (2.8.14) from Proposition 2.8.5, we have

$$c_{\boldsymbol{\mathcal{H}}^{EA}} \|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}} - \widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \leq \mathcal{D}_\infty(\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}}) < \epsilon, \tag{2.4.13}$$

with probability $\mathcal{P}$. With the same probability,

$$\begin{aligned}
\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} &\leq 2\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}} - \widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} + 2\|\widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \\
&\leq \frac{2}{c_{\boldsymbol{\mathcal{H}}^{EA}}}\bigg(\epsilon + \inf_{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}} K^2 \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}\bigg) \\
&\leq \frac{2}{c_{\boldsymbol{\mathcal{H}}^{EA}}}\bigg(\epsilon + \inf_{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}} K^2 \max\{R, R_{\dot{x}}\}^2 \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\infty}\bigg).
\end{aligned}$$

The first inequality follow from the coercivity condition (2.4.5) and the definition of $\widehat{\boldsymbol{\phi}}^{EA}_{\infty}$. The second follows by the definition of the norms. Now for a chosen $0 < \delta < 1$, let

$$1 - \mathcal{N}\bigg(\boldsymbol{\mathcal{H}}^{EA}, \frac{\epsilon}{48 S_{EA} \max\{R, R_{\dot{x}}\}^2 K^4}\bigg) \exp\bigg(- \frac{c_{\boldsymbol{\mathcal{H}}^{EA}} M \epsilon}{1152 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^6}\bigg) \geq 1 - \delta$$

and solve for M. The proof for the part of the system involving ξ is similar. □

Consistency

In the regime where $M \to \infty$, we will choose an increasing sequence of hypothesis spaces, each satisfying the conditions of Theorem 2.4.5. By our assumptions on the interaction kernels, we can also choose the sequence of $\boldsymbol{\mathcal{H}}_M^{EA}$'s such that the approximation error goes to 0 as $M \to \infty$. This enables us to control the infimum on the right hand side of (2.4.11). From here we can apply Theorem 2.4.5 for each M to prove the consistency of our estimators with respect to the $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})$ norm and derive the following consistency theorem.

Theorem 2.4.6 (Strong Consistency)**.** *Suppose that*

$$\{\boldsymbol{\mathcal{H}}_M^{EA}\}_{M=1}^{\infty} \subset \boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R} \times \mathbf{S}^E) \oplus \boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R} \times \mathbf{S}^A)$$

is a family of compact and convex subsets such that the approximation error goes to zero,

$$\inf_{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}_M^{EA}} \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|_{\infty} \xrightarrow{M \to \infty} 0.$$

Further suppose that the coercivity condition holds on $\bigcup_M \boldsymbol{\mathcal{H}}_M^{EA}$*, and that* $\bigcup_M \boldsymbol{\mathcal{H}}_M^{EA}$ *is compact in* $\boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R} \times \mathbf{S}^E) \oplus \boldsymbol{L}^{\boldsymbol{\infty}}(\mathbf{R} \times \mathbf{S}^A)$*. Then the estimator is strongly consistent with respect to the* $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})$ *norm:*

$$\lim_{M \to \infty} \|\widehat{\boldsymbol{\phi}}_M^{EA} - \boldsymbol{\phi}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} = 0 \quad \textit{with probability one.}$$

An analogous consistency result holds for the estimator in the ξ *variable.*

These two results together provide a consistency result on the full estimation of the triple $(\widehat{\boldsymbol{\phi}^{\xi}}, \widehat{\boldsymbol{\phi}^{E}}, \widehat{\boldsymbol{\phi}^{A}})$ and thus consistency of our estimation procedure on the full system (2.2.2).

Proof of Theorem 2.4.6. To simplify the notation, we use the same conventions as the proof of Theorem 2.4.5 and let $\mathcal{D}_\infty = \mathcal{D}_{L,\infty,\boldsymbol{\mathcal{H}}_M^{EA}}$. By definition of the coercivity constant in (2.4.5), we have the inequality $c_{\cup_M \boldsymbol{\mathcal{H}}_M^{EA}} \le c_{\boldsymbol{\mathcal{H}}_M^{EA}}$. From an argument analogous to the one used to arrive at equation (2.4.13) in the proof of Theorem 2.4.5, we obtain that

$$c_{\cup_M \boldsymbol{\mathcal{H}}_M^{EA}} \|\widehat{\boldsymbol{\phi}}_M^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \le \mathcal{D}_\infty(\widehat{\boldsymbol{\phi}}_M^{EA}) + \boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}_\infty^{EA}) \,. \tag{2.4.14}$$

For $\epsilon > 0$, inequality (2.4.14) implies

$$\begin{aligned} P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{c_{\cup_M \boldsymbol{\mathcal{H}}_M^{EA}} \|\widehat{\boldsymbol{\phi}}_M^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \ge \epsilon\} &\le P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\mathcal{D}_\infty(\widehat{\boldsymbol{\phi}}_M^{EA}) + \boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}_\infty^{EA}) \ge \epsilon\} \\ &\le P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\mathcal{D}_\infty(\widehat{\boldsymbol{\phi}}_M^{EA}) \ge \frac{\epsilon}{2}\right\} + P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}_\infty^{EA}) \ge \frac{\epsilon}{2}\right\}. \end{aligned}$$

We now bound the two terms in the above expression separately. For the first term, the proof of Theorem 2.4.5 shows that

$$\begin{aligned} P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\boldsymbol{\mathcal{D}}_\infty(\widehat{\boldsymbol{\phi}}_M^{EA}) \ge \frac{\epsilon}{2}\} &\le \mathcal{N}\left(\boldsymbol{\mathcal{H}}_M^{EA}, \frac{\epsilon}{C_1}\right) \exp\left(-\frac{c_{\boldsymbol{\mathcal{H}}_M^{EA}} M\epsilon}{C_2}\right) \\ &\le \mathcal{N}\left(\cup_M \boldsymbol{\mathcal{H}}_M^{EA}, \frac{\epsilon}{C_1}\right) \exp\left(-\frac{c_{\cup_M \boldsymbol{\mathcal{H}}_M^{EA}} M\epsilon}{C_2}\right) \end{aligned}$$

where $C_1 = 96 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4$, $C_2 = 2304 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4$, and $\mathcal{N}(\cup_M \boldsymbol{\mathcal{H}}_M^{EA}, \frac{\epsilon}{C_1})$ is finite because of the compactness assumption on $\cup_M \boldsymbol{\mathcal{H}}_M^{EA}$. Summing this bound in M,

$$\sum_{M=1}^\infty P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\mathcal{D}_\infty(\widehat{\boldsymbol{\phi}}_M^{EA}) \ge \frac{\epsilon}{2}\} \le \mathcal{N}\left(\cup_M \boldsymbol{\mathcal{H}}_M^{EA}, \frac{\epsilon}{C_1}\right) \sum_{M=1}^\infty \exp\left(-\frac{c_{\cup_M \boldsymbol{\mathcal{H}}_M^{EA}} M\epsilon}{C_2}\right) < \infty \,.$$

For the second term, the bound (2.8.4) yields

$$\boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}_\infty^{EA}) \le 4K^4 S_{EA} \max\{R, R_{\dot{x}}\}^2 \inf_{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}_M^{EA}} \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|_\infty \xrightarrow{M\to\infty} 0.$$

Since ϵ is fixed, the above result, together with our assumption on the sequence of hypothesis spaces, implies that $P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\widehat{\boldsymbol{\phi}}_{\infty}^{EA}) \geq \frac{\epsilon}{2}\right\} = 0$ for M sufficiently large. So we have $\sum_{M=1}^{\infty} P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\widehat{\boldsymbol{\phi}}_{\infty}^{EA}) \geq \frac{\epsilon}{2}\} < \infty$. The finiteness of the two sums above implies, by the first Borel-Cantelli Lemma, that

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\limsup_{M\to\infty}\{c_{\cup_M \boldsymbol{\mathcal{H}}_M^{EA}}\|\widehat{\boldsymbol{\phi}}_M^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \geq \epsilon\}\right\} = 0\,.$$

Since ϵ was arbitrary, we have the desired strong consistency of the estimator. An exactly analogous argument gives the result on the part of the system involving ξ. $\square$

2.4.4 Rate of convergence

Theorem 2.4.5 highlights the classical bias-variance tradeoff in our setting. Given data collected from M trajectories, we would like to choose the best hypothesis space to maximize the accuracy of the estimators. On the one hand, we would like the hypothesis space $\boldsymbol{\mathcal{H}}_M^{EA}$ to be large so that the bias

$$\inf_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}_M^{EA}} \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}\,, \text{ or } \quad \inf_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}} \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\infty}\,,$$

is small. Simultaneously, we would like $\boldsymbol{\mathcal{H}}_M^{EA}$ to be small enough so that the covering number $\mathcal{N}(\boldsymbol{\mathcal{H}}_M^{EA}, \epsilon)$ is small. Just as in nonparametric regression, our rate of convergence depends on a regularity condition of the true interaction kernels and corresponding approximation properties of the hypothesis space, as is demonstrated in the following theorem. We establish the optimal (up to a log factor) min-max rate of convergence by choosing a hypothesis space of an optimal dimension as a function of the sample size M.

The dimension of the space supporting $\boldsymbol{\rho}_T^{EA,L}$ is typically large: it is equal to $1 + \sum_{kk'} p^E_{(k,k')} + \sum_{(k,k')} p^A_{(k,k')}$, see table 2.1 for the definition of the $p_{(k,k')}$. However we can exploit the structure of the system in such a way that our convergence rate

only depends on the maximum number of unique variables in a pair (k, k') across the E, A portions of the system. A similar result holds for the $\boldsymbol{\rho}_T^{\xi,L}$ and its convergence rate. For the system (2.2.2), consider $\mathcal{V}^{E,kk'}$ to be the number of distinct variables in the function $\phi_{kk'}^E(r, \boldsymbol{s}_{(k,k')}^E)$, similarly we define $\mathcal{V}^{A,kk'}$, $\mathcal{V}^{\xi,kk'}$, more precisely, and recalling the notation in table 2.1:

$$\begin{aligned} \mathcal{V}^{E,kk'} &:= 1 + p_{(k,k')}^E \\ \mathcal{V}^{A,kk'} &:= 1 + p_{(k,k')}^A \\ \mathcal{V}^{\xi,kk'} &:= 1 + p_{(k,k')}^\xi \end{aligned} \tag{2.4.15}$$

Using these, we get the dimensions needed for the minimax rates:

$$\mathcal{V} := \max_{k,k'} \{\mathcal{V}^{E,kk'}, \mathcal{V}^{A,kk'}\}, \quad \mathcal{V}^\xi := \max_{k,k'} \mathcal{V}^{\xi,kk'} \tag{2.4.16}$$

The dimension for the minimax rates on the energy and alignment inference is given by $\mathcal{V}$, representing the maximum number of unique variables used in any one of the $\phi_{kk'}^E, \phi_{kk'}^A$ pairs. Analogously, $\mathcal{V}^\xi$ is used for the minimax convergence rate for the inference of $\boldsymbol{\phi}^\xi$. As an extreme example, consider a problem with 10 different types of agents, leading to 100 distinct interaction kernels, each depending on r and one additional variable that is unique for each function. In this case, we only pay the 2 dimensional rate, rather than the 101-dimensional rate in the ambient space of the 101 unique variables, although the heterogeneity affects the constant in the convergence rate. We note that we are not predicting the number of variables nor their form: these are assumed known.

Theorem 2.4.7 (Rate of Convergence)**.** *Let $\widehat{\boldsymbol{\phi}}^{EA} := \widehat{\boldsymbol{\phi}}_M^E \oplus \widehat{\boldsymbol{\phi}}_M^A$ denote the minimizer of the empirical error functional $\boldsymbol{\mathcal{E}}_M^{EA}$ (defined in* (2.3.4)*) over the hypothesis space $\boldsymbol{\mathcal{H}}_M^{EA}$.*

(a) Let the hypothesis space be chosen as the direct sum of the admissible spaces, namely $\boldsymbol{\mathcal{H}}^{EA} = \boldsymbol{\mathcal{K}}^{E}_{S_E} \oplus \boldsymbol{\mathcal{K}}^{A}_{S_A}$*, and assume that the coercivity condition* (2.4.5) *holds on* $\boldsymbol{\mathcal{H}}^{EA}$*. Then, there exists a constant* C *depending only on* $K, S_{EA}, R, R_{\dot{x}}$ *such that*

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\|\widehat{\boldsymbol{\phi}}^{EA}_{M} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{EA,L}_T)}\right] \leq \frac{C}{c_{\boldsymbol{\mathcal{H}}^{EA}}} M^{-\frac{1}{\mathcal{V}+1}}.$$

(b) Assume that $\{\boldsymbol{\mathcal{L}}_n\}_{n=1}^{\infty}$ *is a sequence of finite-dimensional linear subspaces of* $\boldsymbol{L}^{\infty}(\mathbf{R}\times\mathbf{S}^E)\oplus\boldsymbol{L}^{\infty}(\mathbf{R}\times\mathbf{S}^A)$ *satisfying the dimension and approximation constraints*

$$dim(\boldsymbol{\mathcal{L}}_n) \leq c_0 K^2 n^{\mathcal{V}}\,, \qquad \inf_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{L}}_n} \|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|_{\infty} \leq c_1 n^{-s}, \tag{2.4.17}$$

for some fixed constants c_0, c_1 *representing dimension-independent approximation characteristics of the linear subspaces, and* $s > 0$ *related to the regularity of the interaction kernels. The value* n *can be thought of as the number of basis functions along each of the (up to)* $\mathcal{V}$ *axes for each* (k, k')*. Suppose the coercivity condition holds true on the set* $\boldsymbol{\mathcal{L}} := \cup_n \boldsymbol{\mathcal{L}}_n$*, and let* $c^{EA}_{\boldsymbol{\mathcal{L}}}$ *be the coercivity constant of* $\boldsymbol{\mathcal{L}}$*. Define* $\boldsymbol{\mathcal{B}}_n$ *to be the closed ball centered at the origin of radius* $(c_1 + S_{EA})$ *in* $\boldsymbol{\mathcal{L}}_n$*. If we choose the hypothesis space as* $\boldsymbol{\mathcal{H}}_M = \boldsymbol{\mathcal{B}}_{k(M)}$*, where* $k(M) \asymp (\frac{M}{\log M})^{\frac{1}{2s+\mathcal{V}}}$*, then there exists a constant* C *depending on* $K, R, R_{\dot{x}}, S_{EA}, c_0, c_1, s$ *such that we achieve the convergence rate,*

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\|\widehat{\boldsymbol{\phi}}^{EA}_{M} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{EA,L}_T)}\right] \leq \frac{C}{c^{EA}_{\boldsymbol{\mathcal{L}}}} \left(\frac{\log M}{M}\right)^{\frac{2s}{2s+\mathcal{V}}}. \tag{2.4.18}$$

(c) under the corresponding assumptions as in (a), there exists a constant C *depending only on* K, S_{ξ}, R *such that*

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\|\widehat{\boldsymbol{\phi}}^{\xi}_{M} - \boldsymbol{\phi}^{\xi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{\xi,L}_T)}\right] \leq \frac{C}{c_{\boldsymbol{\mathcal{H}}^{\xi}}} M^{-\frac{1}{\mathcal{V}^{\xi}+1}}.$$

(d) under the corresponding assumptions as in (b), there exists a constant C depending only on K, R, S_ξ, c_0, c_1, s such that, and for c^ξ the coercivity constant of the corresponding linear space,

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\|\widehat{\boldsymbol{\phi}}^{\xi}_{M} - \boldsymbol{\phi}^{\xi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{\xi,L}_T)}\right] \leq \frac{C}{c^\xi}\left(\frac{\log M}{M}\right)^{\frac{2s}{2s+\mathcal{V}^\xi}} . \tag{2.4.19}$$

We in fact prove bounds not only in expectation, but also with high probability, for every fixed large-enough M, as the proof will show.

Proof of Theorem 2.4.7. For part (a), let $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{K}}^E_{S_E} \oplus \boldsymbol{\mathcal{K}}^A_{S_A}$. Standard results on covering numbers of function spaces (see theorem 2.7.1 of [134]) give us that the covering number of $\boldsymbol{\mathcal{H}}$ satisfies

$$\mathcal{N}(\boldsymbol{\mathcal{H}}, \epsilon, \|\cdot\|_\infty) \leq C_{\boldsymbol{\mathcal{H}}} \exp\left(\sum_{k,k'=1,1}^{K,K}\left(\frac{1}{\epsilon}\right)^{\mathcal{V}^{E,kk'}} + \left(\frac{1}{\epsilon}\right)^{\mathcal{V}^{A,kk'}}\right) \leq C_{\boldsymbol{\mathcal{H}}} \exp\left(2K^2\left(\frac{1}{\epsilon}\right)^{\mathcal{V}}\right)$$

for some absolute constant $C_{\boldsymbol{\mathcal{H}}}$ depending only on $\boldsymbol{\mathcal{H}}$ and $\mathcal{V}$. By assumption on the hypothesis space, we have that $\inf_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}}\|\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}\|^2_\infty = 0$. From this, the concentration estimate (2.4.11) together with the covering number bound imply that,

$$\begin{aligned} P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}^{EA,L}_T)} > \epsilon\} &\leq \mathcal{N}(\boldsymbol{\mathcal{H}}, C_1\epsilon, \|\cdot\|_\infty)\exp(-C_2M\epsilon) \\ &\leq C_{\boldsymbol{\mathcal{H}}}\exp(2K^2(C_1\epsilon)^{-\mathcal{V}} - C_2M\epsilon) \end{aligned} \tag{2.4.20}$$

where $C_1 = \frac{c_{\boldsymbol{\mathcal{H}}}}{48S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}$ and $C_2 = \frac{c_{\boldsymbol{\mathcal{H}}}}{1152S^2_{EA}\max\{R,R_{\dot{x}}\}^2K^4}$. Next, define the function

$$g(\epsilon) := 2K^2(C_1\epsilon)^{-\mathcal{V}} - \frac{C_2M\epsilon}{2},$$

which we will minimize to achieve the desired probability bound. By direct calculation, $g(\epsilon) = 0$ if we choose $\epsilon = \epsilon_M = (\frac{C_3}{M})^{\frac{1}{\mathcal{V}+1}}$, where $C_3 = \left(\frac{4K^2}{C_2 C_1^{\mathcal{V}}}\right)^{\frac{1}{\mathcal{V}+1}}$; moreover the derivative of $g(\epsilon)$ is ≤ 0 for all $\epsilon \geq \epsilon_M$. Therefore, the bound (2.4.20) implies

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > \epsilon\} \leq \begin{cases} \exp\left(\frac{-C_2 M\epsilon}{2}\right), & \epsilon \geq \epsilon_M \\ 1, & \epsilon \leq \epsilon_M \end{cases} \tag{2.4.21}$$

Integrating over $\epsilon \in (0, +\infty)$ and using $e^{-x} \leq x + 1$ for all $x \geq 0$, we obtain

$$\int_0^\infty P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > \epsilon\} d\epsilon \leq \Big(\frac{C_4}{M}\Big)^{\frac{1}{\mathcal{V}+1}} + O\Big(\frac{1}{M}\Big)$$

Using coercivity and (2.4.11), we conclude that

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}[\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}^{EA}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}] \leq \frac{C_4}{c_{\boldsymbol{\mathcal{H}}^{EA}}} M^{-\frac{1}{\mathcal{V}+1}},$$

where C_4 is an absolute constant that only depends on $K, S_{EA}, R, R_{\dot{x}}$.

For part (b), we recall (see [45, Proposition 5]) that

$$\mathcal{N}(\boldsymbol{\mathcal{B}}_n, \epsilon, \|\cdot\|_\infty) \leq \left(\frac{4(c_1 + S_{EA})}{\epsilon}\right)^{c_0 K^2 n^{\mathcal{V}}}.$$

Using (2.4.11), and the approximation assumption, we bound the probability as

$$\begin{aligned}
&P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{B}}_n} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > \epsilon + c_2 n^{-2s}\} \\
&\quad = P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{B}}_n} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > t' n^{-2s} + c_2 n^{-2s}\} \\
&\quad = P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{B}}_n} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > t n^{-2s}\} \\
&\quad \leq \mathcal{N}\Big(\boldsymbol{\mathcal{B}}_n, c_3' t n^{-2s}, \|\cdot\|_\infty\Big) \exp(-c_4 M t n^{-2s}) \\
&\quad \leq \Big(\frac{c_3}{t n^{-2s}}\Big)^{c_0 K^2 n^{\mathcal{V}}} \exp(-c_4 M t n^{-2s}) \\
&\quad \leq \exp(c_0 K^2 n^{\mathcal{V}} \log(c_3) + c_0 K^2 n^{\mathcal{V}} |\log(t n^{-2s})| - c_4 M t n^{-2s}),
\end{aligned} \tag{2.4.22}$$

where $c_2 = \frac{1}{c_{\cup_n} \boldsymbol{\mathcal{L}}_n} c_1$, $c_3' = \frac{c_{\cup_n} \boldsymbol{\mathcal{L}}_n}{48(S_{EA}+c_1)\max\{R,R_{\dot{x}}\}^2 K^4}$, $c_3 = \frac{192(S_{EA}+c_1)^2 \max\{R,R_{\dot{x}}\}^2 K^4}{c_{\boldsymbol{\mathcal{L}}}^{EA}}$, and $c_4 = \frac{c_{\cup_n} \boldsymbol{\mathcal{L}}_n}{1152(S_{EA}+c_1)^2 \max\{R,R_{\dot{x}}\}^2 K^4}$ are absolute constants independent of M. Define

$$g(n) := c_0 n^{\mathcal{V}} K^2 \log(c_3) + c_0 n^{\mathcal{V}} K^2 |\log(t n^{-2s})| - \frac{c_4}{2} M t n^{-2s}.$$

To find the optimal n in terms of M, we minimize g in n. By taking a derivative, and solving the corresponding equation, we see that the optimal n is

$$n_* = O\bigg(\Big(\frac{M}{\log M} \Big)^{\frac{1}{2s+\mathcal{V}}} \bigg),$$

with a constant depending on c_3, c_4, c_2 but not on M. We choose $n_* = \lfloor (\frac{M}{\log M})^{\frac{1}{2s+\mathcal{V}}} \rfloor$, let $\epsilon_M = (\frac{M}{\log M})^{\frac{2s}{2s+\mathcal{V}}}$ and

$$h(\epsilon) := c_0 n_* K^2 \log(c_3) + c_0 n_* K^2 |\log(\epsilon)| - \frac{c_4}{2} M \epsilon.$$

As above, let $\epsilon = t n_*^{-2s} = t\epsilon_M$ and consider $h(t\epsilon_M)$. It is easy to see that $\lim_{t \to 0^+} h(t\epsilon_M) = \infty$ and $\lim_{t \to \infty} h(t\epsilon_M) = -\infty$. Together with the continuity of h, these facts imply that there exists a constant c_5, depending on K, c_0, c_2, c_3, c_4, such that $h(c_5 \epsilon_M) = 0$. We further need that $h'(\epsilon) \le 0$ for all $\epsilon \ge c_5 \epsilon_M$. By taking the derivative of h, setting it ≤ 0, we find that this condition eventually holds by basic calculus. Therefore, if needed, to satisfy the derivative condition, we can enlarge the constant c_5 to a constant c_6 (independent of M) such that $h(\epsilon) \le 0$ and $h'(\epsilon) \le 0$ for all $\epsilon \ge c_6 \epsilon_M$. These results imply

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \{ \| \widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{B}}_{n_*}} - \boldsymbol{\phi}^{EA} \|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > \epsilon \} \le \begin{cases} \exp\left(\frac{-c_4 M \epsilon}{2} \right), & \epsilon \ge c_6 \epsilon_M \\ 1, & \epsilon \le c_6 \epsilon_M \end{cases}, \tag{2.4.23}$$

and therefore

$$\int_0^\infty P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\{\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{B}}_{n_*}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} > \epsilon\}d\epsilon \leq C_1\Big(\frac{\log M}{M}\Big)^{\frac{2s}{2s+\mathcal{V}}},$$

where C_1 is a constant depending on $c_0, c_1, s, K, S_{EA}, R, R_{\dot{x}}$.

Now, with $\boldsymbol{\mathcal{H}}_M^{EA} = \boldsymbol{\mathcal{B}}_{n_*}$ and using (2.4.11), we have shown the convergence rate

$$\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}[\|\widehat{\boldsymbol{\phi}}^{EA}_{L,M,\boldsymbol{\mathcal{H}}_M^{EA}} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}] \leq \frac{c_7}{c_{\boldsymbol{\mathcal{L}}}^{EA}}\Big(\frac{M}{\log M}\Big)^{-\frac{2s}{2s+\mathcal{V}}},$$

where c_7 is an absolute constant that only depends on $s, K, c_0, c_1, S_{EA}, R, R_{\dot{x}}$.

□

In both theorems, the convergence rates $\frac{2s}{2s+\mathcal{V}}$ and $\frac{2s}{2s+\mathcal{V}^\xi}$ coincide with the minimax rate of convergence $\frac{2s}{2s+d}$ for nonparametric regression in the corresponding dimension d – up to the logarithmic factor. (This logarithmic factor may be removable (using, e.g., the techniques in Chapter 11-15 of [64]), but with additional complexity of the proofs.) Achieving the same rate of convergence as if we had observed the noisy values of the interaction kernels directly, rather than through the dynamics, demonstrates the optimality of our approach. The strong consistency results show the asymptotic optimality of our method, and for wide classes of systems the assumptions of the theorems apply. Specifically, for part (b) of the theorems, the dimension and approximation conditions can be explicitly achieved by piecewise polynomials or splines appropriately adapted to the regularity of the interaction kernel. In the conditions of theorem 2.4.7, n can be the number of partitions along each axis of the variables in $\mathcal{V}$. Then, using multivariate splines or piecewise polynomials we will have a fixed constant c_0 (corresponding to the number of parameters to estimate for each function) times $Kn^{\mathcal{V}}$ as the dimension of the linear space. Furthermore, by standard approximation theory results, see [120] (Chapters 12,13), [51],[53], for s the regularity of the interaction kernels we achieve the desired approximation condition

with piecewise polynomials of degree $\lfloor s \rfloor$. In our admissible spaces we have $s = 1$, note that the rate of convergence is faster if we have a kernel of higher regularity.

We next briefly examine the convergence rate on a few systems of interest. Recall that in table 2.2 we have as the final two columns the values $\mathcal{V}, \mathcal{V}^{\xi}$, which dictate the rate of convergence of our estimators in each system. Some specific highlights:

- For Anticipation Dynamics (AD), even though we are learning both an energy and alignment kernel, because there are only 2 unique variables shared across both of them we learn at the 2-dimensional rate.

- For the Synchronized Oscillator we achieve the 2-dimensional optimal learning rate on each of the EA and ξ portions (rather than a 4-dimensional rate) due to the decoupled nature of the system; similarly we only pay the 1-dimensional rate twice for the Phototaxis system. This is a key reason for splitting our learning theory between EA- and ξ-interaction kernels and accounting for shared and non-shared variables.

- Due to the design of the measures, norms and the associated learning algorithm, even in the heterogeneous case for celestial mechanics and predator-swarm, we only pay the 1-dimensional learning rate, although the constants are of course affected by the heterogeneity and the algorithm requires a larger learning matrix.

- The rates of convergence of our estimators for all previously-studied first-order systems (see [89, 90, 146]) can be derived from Theorem 2.4.7.

One downside of the results above is the lack of dependence on L: it seems natural that finer time samples in each trajectory should improve the results, at least up to a point. Indeed, the numerical experiments of [146, 89, 90] demonstrate that more data in L may indeed be helpful to improve the performance. One technique used in [146] for very long trajectory data (large L, medium to small M) is to split each trajectory

into larger M with smaller L in each. The dependence on the number of agents N is not the objective of this work; it was considered [19] in the case of first-order systems; but further study in this mean-field regime is of interest to the authors and work is ongoing.

2.4.5 Performance of trajectory prediction

Once estimators $\widehat{\boldsymbol{\varphi}}^{EA}, \widehat{\boldsymbol{\varphi}}^{\xi}$ are obtained, a natural question is the accuracy of the evolved trajectory based on these estimated kernels. We compare the observed trajectories to the estimated trajectories evolved from the same initial conditions but with the estimated interaction kernels. Recall that $\boldsymbol{Y}_t = [\boldsymbol{X}_t^T, \boldsymbol{V}_t^T, \boldsymbol{\Xi}_t^T]^T$ be the trajectory from dynamics generated by the true and unknown interaction kernels with initial condition $\boldsymbol{Y}_0$, and $\hat{\boldsymbol{Y}}_t = [\hat{\boldsymbol{X}}_t^T, \hat{\boldsymbol{V}}_t^T, \hat{\boldsymbol{\Xi}}_t^T]^T$ be the trajectory from dynamics generated, with the same initial condition $\hat{\boldsymbol{Y}}_0 = \boldsymbol{Y}_0$, by the interaction kernels estimated from observations at times $\{t_l\}_{l=1}^L$. We let

$$\left\|\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_t\right\|_{\mathcal{Y}}^2 := \left\|\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t\right\|_{\mathcal{S}}^2 + \left\|\boldsymbol{V}_t - \hat{\boldsymbol{V}}_t\right\|_{\mathcal{S}}^2 + \left\|\boldsymbol{\Xi}_t - \hat{\boldsymbol{\Xi}}_t\right\|_{\mathcal{S}}^2. \tag{2.4.24}$$

The next theorem shows that the error in prediction is (i) bounded trajectory-wise by a continuous-time version of the error functional, and (ii) bounded on average by the $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA}), \boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi})$, respectively, error of the estimator. This further validates the effectiveness of our error functional and $\boldsymbol{L}^2(\boldsymbol{\rho}_T)$-metric to assess the quality of the estimator. In particular, this emphasizes that although the system is a coupled system of ODE's, our decoupled learning procedure with our choice of norm will lead to control of the expected supremum error as long as we minimize the $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA}), \boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi})$ norms in obtaining our estimators.

Theorem 2.4.8. *Suppose that $\widehat{\boldsymbol{\phi}}^E \in \boldsymbol{\mathcal{K}}_{S_E}^E$, $\widehat{\boldsymbol{\phi}}^A \in \boldsymbol{\mathcal{K}}_{S_A}^A$ and $\widehat{\boldsymbol{\phi}}^{\xi} \in \boldsymbol{\mathcal{K}}_{S_\xi}^{\xi}$. Denote by $\widehat{\boldsymbol{Y}}(t)$ and $\boldsymbol{Y}(t)$ the solutions of the systems with kernels $\widehat{\boldsymbol{\phi}}^E = (\widehat{\phi}_{kk'}^E)_{k,k'=1}^{K,K}, \widehat{\boldsymbol{\phi}}^A = (\widehat{\phi}_{kk'}^A)_{k,k'=1}^{K,K}$,*

and $\widehat{\boldsymbol{\phi}}^{\xi} = (\widehat{\phi}^{\xi}_{kk'})_{k,k'=1}^{K,K}$ *and* $\boldsymbol{\phi}^{E}, \boldsymbol{\phi}^{A}, \boldsymbol{\phi}^{\xi}$ *respectively, both with the same initial condition. Then*

$$\begin{aligned}\sup_{t\in[0,T]} \|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2 \le g(T)\Bigg[2T^2 \int_{u=0}^{t}\int_{s=0}^{u} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{nc,\dot{\boldsymbol{x}}}(\boldsymbol{X}_s,\boldsymbol{V}_s,\boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\boldsymbol{\phi}}^{EA}}(\boldsymbol{X}_s,\boldsymbol{V}_s,\boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 \, ds\, du \\ + 2T\int_{s=0}^{t} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{nc,\dot{\boldsymbol{x}}}(\boldsymbol{X}_s,\boldsymbol{V}_s,\boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\boldsymbol{\phi}}^{EA}}(\boldsymbol{X}_s,\boldsymbol{V}_s,\boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds \\ + 2T\int_{s=0}^{t} \|\dot{\boldsymbol{\Xi}}_s - \mathbf{f}^{nc,\xi}(\boldsymbol{X}_s,\boldsymbol{V}_s,\boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\boldsymbol{\phi}}_{\xi}}(\boldsymbol{X}_s,\boldsymbol{V}_s,\boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds\Bigg]\end{aligned}$$

where $g(T) = 1 + (1 + B_1 T)T\exp(A_1 T + T^2/2)$*. The constants are* $A_1 = 2T(8KP + \mathcal{L} + 8QK + \mathcal{L}^{\xi})$ *and* $B_1 = 2T^2(8KP + \mathcal{L})$*, with any unspecified constants made precise in the proof and only depending on the Lipschitz constants of the noncollective forces and the feature maps, as well as the values* S^E, S^A, S^{ξ} *coming from the admissible spaces. It is bounded on average, with respect to the initial distribution* $\boldsymbol{\mu}^{\boldsymbol{Y}}$*, by*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}[\sup_{t\in[0,T]} \|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2] \le g(T)\Bigg((T^2K^2 + TK^2)\|\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA})} \\ + TK^2\|\widehat{\boldsymbol{\phi}}^{\xi} - \boldsymbol{\phi}^{\xi}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi})}\Bigg)\end{aligned} \tag{2.4.25}$$

with the measures $\boldsymbol{\rho}_T^{EA}, \boldsymbol{\rho}_T^{\xi}$ *defined in (2.4.1, 2.13.1). Expression (2.4.25) shows that by minimizing the right hand side, we can control the expected* $\mathcal{Y}$*-supremum error of the estimated trajectories.*

We postpone the somewhat lengthy proof to Appendix 2.7.

2.5 Applications

Our learning theory, as well as measures, norms, functionals etc. can be applied to study all the examples considered in the works [89, 90, 146]. These examples, particularly those of [146], can thus be considered as applications of the theoretical results as well as of the algorithm in section 2.14. We choose to study two new

dynamics, which are not considered in [89, 146] since they exhibit some unique features of our generalized model. In particular, we choose them due to their special form of having both energy-based and alignment-based interactions. These are the flocking with external potential (FwEP) model in [122] and the anticipation dynamics (AD) model in [121].

Table 2.4 shows the value of learning parameters for these dynamics.

M_ρ	L	T_f	T	$\mu^{\boldsymbol{x}}$	$\mu^{\dot{\boldsymbol{x}}}$	Num. of learning trials
2000	500	10	5	Unif.$([0,5]^2)$	Unif.$([0,5]^2)$	10

Table 2.4: Values of parameters for the learning.

The setup of the learning experiment is as follows. We use M_ρ different initial conditions to evolve the dynamics[1] from 0 to T for the sole purpose of obtaining a good approximation to $\rho_T^{L,EA}, \rho_T^{L,E}$ and $\rho_T^{L,A}$. Then we use another set of M ($M = 500$ for FwEP and $M = 750$ for AD) initial conditions to generate training data to learn the corresponding ϕ^E and ϕ^A from the empirical distributions, $\rho_T^{L,M,EA}$'s, etc. We report the relative learning errors calculated via (2.4.4) for $\hat{\phi}^E \oplus \hat{\phi}^A$, (2.4.4) for $\hat{\phi}^E$, and (2.4.4) for $\hat{\phi}^A$, along with pictorial comparison of those interaction kernels as well as a visualization on the pairwise data which is used to learn the estimated kernels. Then we evolve the dynamics either from the training set of M initial conditions or another set of M randomly chosen initial conditions with $\phi^E \oplus \phi^A$ and $\hat{\phi}^E \oplus \hat{\phi}^A$ from 0 to $T_f > T$, and report the trajectory errors calculated using (2.5.1) on $\boldsymbol{y}$ (the whole system), and for $\boldsymbol{x}$ (the position) and $\boldsymbol{v}$ (the velocity). Again, pictorial comparison of the trajectories are also shown. We report the trajectory errors over $[0, T]$ and $[T, T_f]$. The learning results are shown in the following sections. We consider a related norm on the trajectory $\boldsymbol{Y}_{[0,T]} = \{\boldsymbol{Y}_{t_l}\}_{l=1}^L$ ($0 = t_1 < \cdots < t_L = T$):

$$\left\| \boldsymbol{Y}_{[0,T]} - \hat{\boldsymbol{Y}}_{[0,T]} \right\|_{\text{traj}} := \max_{l=1,\ldots,L} \left\| \boldsymbol{Y}(t_l) - \hat{\boldsymbol{Y}}(t_l) \right\|_{\mathcal{Y}} . \tag{2.5.1}$$

[1] We use the built-in MATLAB integrating routine, *ode*15*s*, with relative tolerance at 10^{-8} and absolute tolerance at 10^{-11}.

We also consider a relative version, invariant under changes of units of measure:

$$\left\| \boldsymbol{Y}_{[0,T]} - \hat{\boldsymbol{Y}}_{[0,T]} \right\|_{\mathrm{traj}^*} := \frac{\left\| \boldsymbol{Y}_{[0,T]} - \hat{\boldsymbol{Y}}_{[0,T]} \right\|_{\mathrm{traj}}}{\left\| \boldsymbol{Y}_{[0,T]} \right\|_{\mathrm{traj}}} .$$

Lastly, we report errors between $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$,

$$\left\| \boldsymbol{X}_{[0,T]} - \hat{\boldsymbol{X}}_{[0,T]} \right\|_{\mathcal{S}^*} := \frac{\max_{l=1,\dots,L} \{ \left\| \boldsymbol{X}(t_l) - \hat{\boldsymbol{X}}(t_l) \right\|_{\mathcal{S}} \}}{\max_{l=1,\dots,L} \{ \| \boldsymbol{X}(t_l) \|_{\mathcal{S}} \}}.$$

Similar re-scaled norms are used for the difference between $\boldsymbol{V}_{[0,T]}$ and $\hat{\boldsymbol{V}}_{[0,T]}$, and for the difference between $\boldsymbol{\Xi}_{[0,T]}$ and $\hat{\boldsymbol{\Xi}}_{[0,T]}$.

2.5.1 Learning results for flocking with external potential

We consider the FwEP model for its simplicity and clustering behavior in both position and velocity (hence flocking occurs). The dynamics of the FwEP model is given as follows,

$$\ddot{\boldsymbol{x}}_i = \frac{1}{N} \sum_{i'=1, i' \neq i}^{N} a(\dot{\boldsymbol{x}}_{i'} - \dot{\boldsymbol{x}}_i) + \frac{1}{N} \sum_{i'=1, i' \neq i}^{N} \phi(\| \boldsymbol{x}_{i'} - \boldsymbol{x}_i \|)(\dot{\boldsymbol{x}}_{i'} - \dot{\boldsymbol{x}}_i).$$

Here $a > 0$ is a constant representing an attraction force, and $\phi = \frac{1}{(1+r^2)^\beta}$[2] with $\beta = \frac{1}{2}$. To fit into our learning regime, we take, $m_i = 1$, $K = 1$, no ξ_i, no non-collective force, and

$$\phi^E := a \quad \text{and} \quad \phi^A := \frac{1}{(1+r^2)^\beta}.$$

For the FwEP model, we use the space of 1^{st} degree piecewise polynomials with dimension $n^E = 122$ for learning ϕ^E; and for ϕ^A, we use the same space. First, consider the comparison of energy-based interactions shown in Fig. 2.1.

[2]This choice of interaction for alignment is not mandatory. It is a good comparison between this FwEP model with the Cucker-Smale mode with this choice of interaction functions.

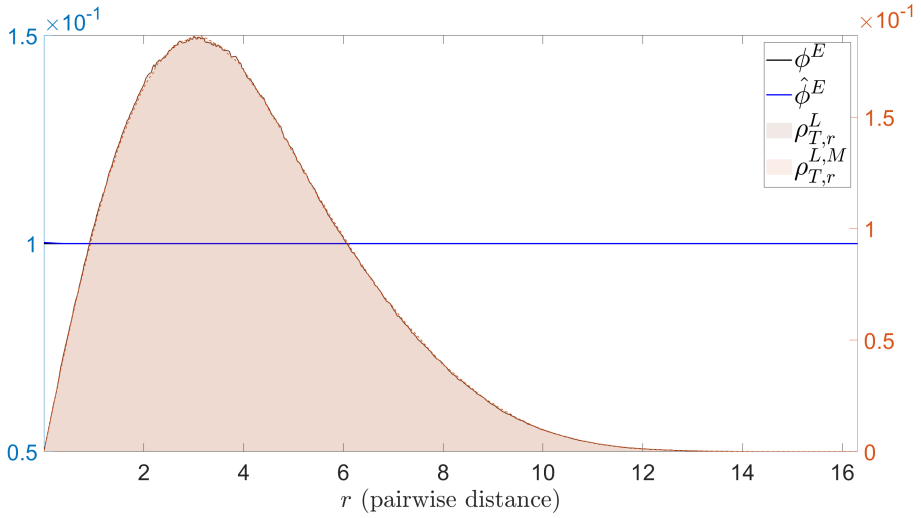
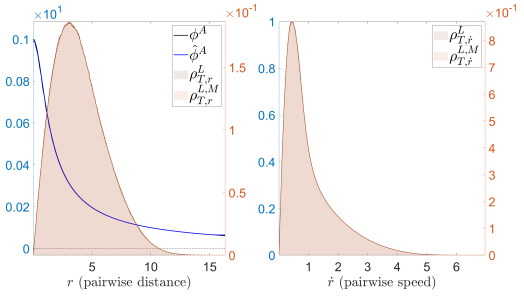
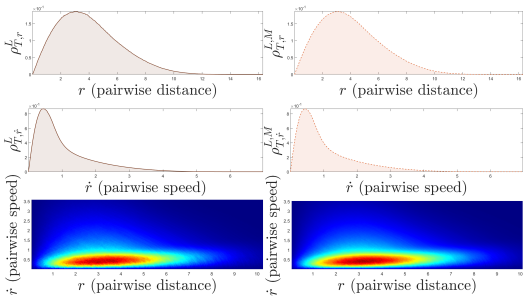


Figure 2.1: ϕ^E vs. $\hat{\phi}^E$, Err: $3.9 \cdot 10^{-6} \pm 6.6 \cdot 10^{-7}$. The lines shown in blue are the estimated interaction kernels, and the lines shown in black are the true interaction kernels. The colored areas shown in the background are the learned distributions of pairwise distance data.

Fig. 2.1 shows that our learning performance on constant functions using piecewise linear polynomials shows promising results. However, we still have trouble learning the behavior of the interaction at $r = 0$, part of it due to the weight of $\vec{0}$ in the model, and the other part of it being lack of available data towards $r = 0$. Next, we show the comparison of alignment-based interactions in Fig. 2.2 with distribution of the pairwise data.



((a)) ϕ^A vs. $\hat{\phi}^A$, Err: $9.1 \cdot 10^{-3} \pm 2.5 \cdot 10^{-4}$.



((b)) $\rho_T^{A,L}$ vs. $\rho_T^{A,L,M}$.

Figure 2.2: The lines shown in blue are the estimated interaction kernels, and the lines shown in black are the true interaction kernels. The colored areas shown in the background are the learned distributions of pairwise distance data.

Again, in Fig. 2.2(a), it shows a faithful approximation from our estimated kernels.

The $\hat{\phi}^E \oplus \hat{\phi}^A$ error is: $5.8 \cdot 10^{-3} \pm 1.6 \cdot 10^{-4}$. The comparison of trajectories are shown in Fig. 2.3.

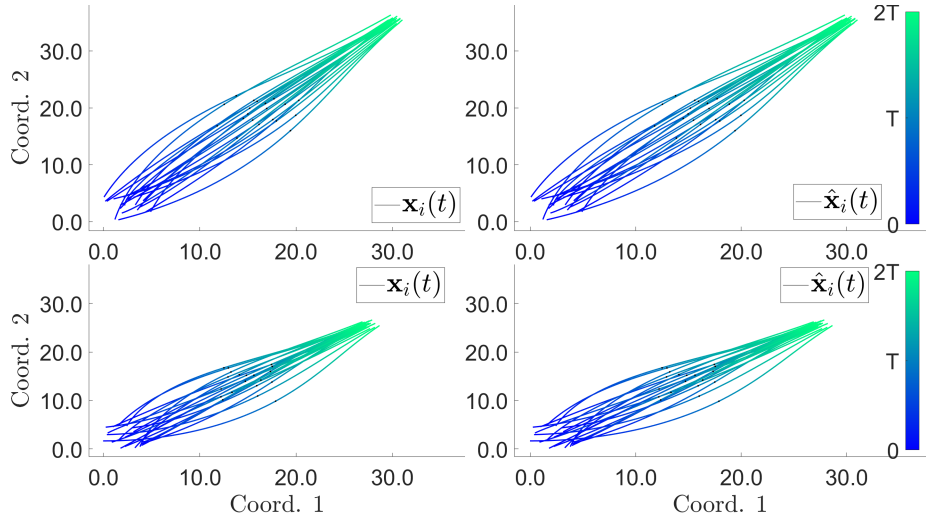


Figure 2.3: Flocking with external potential trajectory comparison.

Fig. 2.3 shows little visual difference between the learned and observed trajectories. A more quantitative description of the trajectory errors are shown in table 2.5.

	$[0,T]$	$[T,T_f]$
mean_{IC} on $\boldsymbol{x}$	$7.2 \cdot 10^{-4} \pm 1.9 \cdot 10^{-5}$	$6.7 \cdot 10^{-4} \pm 1.8 \cdot 10^{-5}$
mean_{IC} on $\boldsymbol{v}$	$1.15 \cdot 10^{-3} \pm 3.1 \cdot 10^{-5}$	$1.5 \cdot 10^{-3} \pm 4.1 \cdot 10^{-3}$
mean_{IC} on $\boldsymbol{y}$	$6.2 \cdot 10^{-6} \pm 1.7 \cdot 10^{-7}$	$2.22 \cdot 10^{-6} \pm 6.8 \cdot 10^{-8}$
std_{IC} on $\boldsymbol{x}$	$1.28 \cdot 10^{-4} \pm 4.2 \cdot 10^{-6}$	$1.22 \cdot 10^{-4} \pm 4.0 \cdot 10^{-6}$
std_{IC} on $\boldsymbol{v}$	$2.20 \cdot 10^{-4} \pm 7.0 \cdot 10^{-6}$	$2.5 \cdot 10^{-4} \pm 1.0 \cdot 10^{-5}$
std_{IC} on $\boldsymbol{y}$	$1.52 \cdot 10^{-6} \pm 5.9 \cdot 10^{-8}$	$6.0 \cdot 10^{-7} \pm 2.6 \cdot 10^{-8}$
mean_{IC} on $\boldsymbol{x}$	$7.2 \cdot 10^{-4} \pm 1.7 \cdot 10^{-5}$	$6.7 \cdot 10^{-4} \pm 1.6 \cdot 10^{-5}$
mean_{IC} on $\boldsymbol{v}$	$1.15 \cdot 10^{-3} \pm 2.7 \cdot 10^{-5}$	$1.46 \cdot 10^{-3} \pm 3.3 \cdot 10^{-5}$
mean_{IC} on $\boldsymbol{y}$	$6.2 \cdot 10^{-6} \pm 1.6 \cdot 10^{-7}$	$2.22 \cdot 10^{-6} \pm 5.4 \cdot 10^{-8}$
std_{IC} on $\boldsymbol{x}$	$1.30 \cdot 10^{-4} \pm 5.8 \cdot 10^{-6}$	$1.24 \cdot 10^{-4} \pm 5.3 \cdot 10^{-6}$
std_{IC} on $\boldsymbol{v}$	$2.25 \cdot 10^{-4} \pm 9.9 \cdot 10^{-6}$	$2.56 \cdot 10^{-4} \pm 9.1 \cdot 10^{-6}$
std_{IC} on $\boldsymbol{y}$	$1.6 \cdot 10^{-6} \pm 7.6 \cdot 10^{-8}$	$6.2 \cdot 10^{-7} \pm 2.4 \cdot 10^{-8}$

Table 2.5: Trajectory Errors. The first three rows of mean trajectory errors are from the training set of initial conditions. The next three rows are standard deviation of the trajectory errors from the training set of initial conditions. The following three rows are mean trajectory errors from a new set of initial conditions. Finally, the last three rows report the standard deviation of the trajectory errors from a new set of initial conditions.

We are maintaining a relative four-digit accuracy in estimating the position, and a

relative three-digit accuracy in estimating the velocity of the agents in the system. Although we are able to reconstruct ϕ^E with a 6-digit accuracy, we are not able to do the same for ϕ^A. The error in $\hat{\phi}^E \oplus \hat{\phi}^A$ reflects this discrepancy by considering the two functions together.

2.5.2 Learning results for anticipation dynamics with $U(r) = \frac{r^p}{p}$

The energy-based interactions are constants in the FwEP models, if we want to consider more complicated models, i.e., interactions depending on pairwise distance and more, the AD models are suitable candidates. The dynamics of the AD model is given as follows,

$$\begin{aligned}\ddot{\boldsymbol{x}}_i = &\frac{1}{N}\sum_{i'=1,i'\neq i}^{N}\frac{\tau U'(\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|)}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|}(\dot{\boldsymbol{x}}_{i'}-\dot{\boldsymbol{x}}_i)\\ &+\frac{1}{N}\sum_{i'=1,i'\neq i}^{N}\Big\{\frac{-\tau U'(\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|)(\boldsymbol{x}_{i'}-\boldsymbol{x}_i)\cdot(\dot{\boldsymbol{x}}_{i'}-\dot{\boldsymbol{x}}_i)}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|^3}\\ &+\frac{\tau U''(\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|)(\boldsymbol{x}_{i'}-\boldsymbol{x}_i)\cdot(\dot{\boldsymbol{x}}_{i'}-\dot{\boldsymbol{x}}_i)}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|^2}+\frac{U'(\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|)}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|}\Big\}(\boldsymbol{x}_{i'}-\boldsymbol{x}_i). \quad (2.5.2)\end{aligned}$$

Here τ measures the amount (in time) of anticipation. In order to fit the model into our learning regime, we take

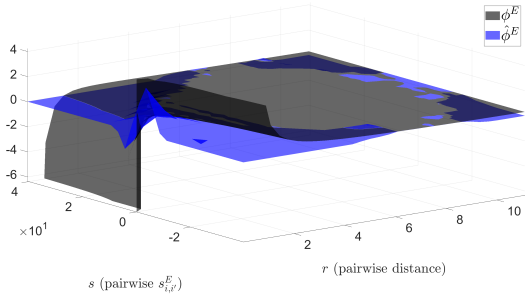
$$\phi^A(r) \coloneqq \frac{\tau U'(r)}{r} \quad \text{and} \quad \phi^E(r,s) \coloneqq \frac{-\tau U'(r)s}{r^3}+\frac{\tau U''(r)s}{r^2}+\frac{U'(r)}{r}.$$

Here we have no ξ_i, $K=1$, $m_i=1$, and

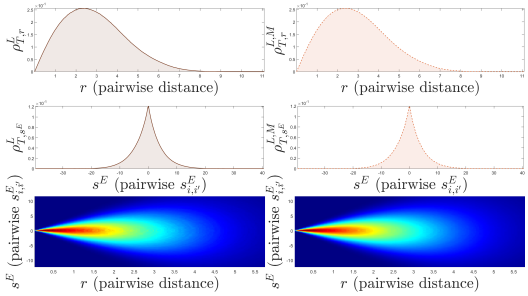
$$s^E_{i,i'} = s^A_{i,i'} \coloneqq (\boldsymbol{x}_{i'}-\boldsymbol{x}_i)\cdot(\dot{\boldsymbol{x}}_{i'}-\dot{\boldsymbol{x}}_i).$$

We also use $\tau = 0.1$.

It is shown in [121] that if U'' is bounded when $r \to \infty$ with $U(0) = U'(0) = 0$, then unconditionally flocking would occur. We take $U(r) = \frac{r^p}{p}$ for $1 < p \leq 2$, then the system would show unconditional flocking. We choose $p = 1.5$ for our learning trials[3]. We use a tensor grid of 1^{st} degree piecewise standard polynomials with $n^E = 28^2$ for learning $\phi^E(r, s)$, then a set of 1^{st} degree piecewise standard polynomials with $n^A = 138$ for learning $\phi^A(r)$. For the energy-based interactions we have the following results.



((a)) $U(r) = \frac{r^{1.5}}{1.5}$: ϕ^E vs. $\hat{\phi}^E$, Err: $6 \cdot 10^{-1} \pm 2.6 \cdot 10^{-1}$.

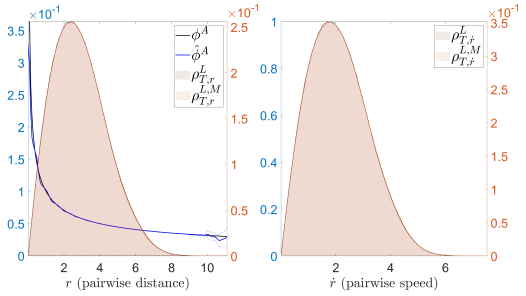


((b)) $U(r) = \frac{r^{1.5}}{1.5}$: $\rho_T^{E,L}$ vs. $\rho_T^{E,L,M}$.

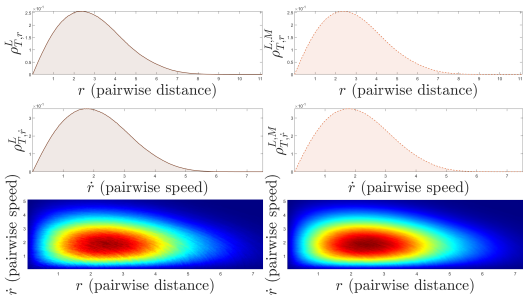
Figure 2.4: The lines shown in blue are the estimated interaction kernels, and the lines shown in black are the true interaction kernels. The colored areas shown in the background are the learned distributions of pairwise distance data.

As is shown in Fig. 2.4(b), the concentration of pairwise distance data is away from 0, making the estimation of the behavior of $\phi^E(r, s)$ at r close to 0 extremely difficult, meanwhile, since ϕ^E is also weighted by the pairwise difference, $\boldsymbol{x}_{i'} - \boldsymbol{x}_i$, and at $r_{i,i'}$ close to 0, the information is also lost. Next, we present the alignment-based interaction kernels in Fig. 2.5(a).

[3] $p = 2$ induces constant forces on the dynamics.



((a)) $U(r) = \frac{r^{1.5}}{1.5}$: ϕ^A vs. $\hat{\phi}^A$. Err: $1.7 \cdot 10^{-1} \cdot 3.9 \cdot 10^{-2}$.



((b)) $U(r) = \frac{r^{1.5}}{1.5}$: $\rho_T^{A,L}$ vs. $\rho_T^{A,L,M}$.

Figure 2.5: The lines shown in blue are the estimated interaction kernels, and the lines shown in black are the true interaction kernels. The colored areas shown in the background are the learned distributions of pairwise distance data.

As shown in Fig. 2.5, the behavior of ϕ^A at $r = 0$ is learned accurately. Less accurate is the estimation of ϕ^A for large r: since the agents have aligned their velocities, the weight $\boldsymbol{v}_{i'} - \boldsymbol{v}_i$ is close to a zero vector. The overall learning performance for estimating ϕ^A is better compared to that of estimating ϕ^E. The $\hat{\phi}^E \oplus \hat{\phi}^A$ error is: $6 \cdot 10^{-1} \pm 3.0 \cdot 10^{-1}$. The comparison of trajectories between the true kernels (LHS) and the estimators (RHS) is shown in Fig.2.6.

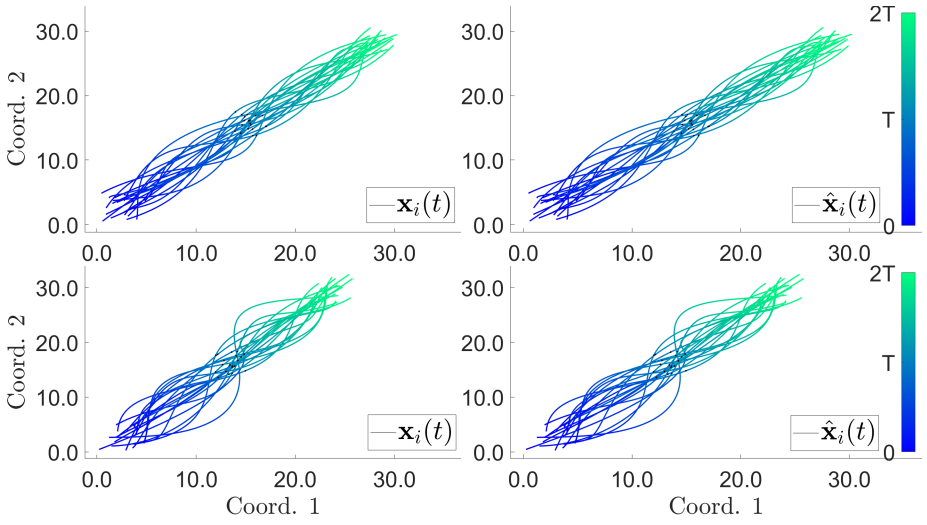


Figure 2.6: $U(r) = \frac{r^{1.5}}{1.5}$: Trajectory Comparison.

As shown in Fig. 2.6, visually, there is no difference between the true dynamics and the estimated dynamics. We offer more quantitative insight into the difference

between the two in table 2.6.

	$[0,T]$	$[T,T_f]$
mean_{IC} on $\boldsymbol{x}$	$2.22 \cdot 10^{-3} \pm 8.5 \cdot 10^{-5}$	$2.4 \cdot 10^{-3} \pm 1.0 \cdot 10^{-4}$
mean_{IC} on $\boldsymbol{v}$	$7.8 \cdot 10^{-3} \pm 3.9 \cdot 10^{-4}$	$1.78 \cdot 10^{-2} \pm 8.9 \cdot 10^{-4}$
mean_{IC} on $\boldsymbol{y}$	$1.91 \cdot 10^{-5} \pm 9.0 \cdot 10^{-7}$	$6.2 \cdot 10^{-6} \pm 3.6 \cdot 10^{-7}$
std_{IC} on $\boldsymbol{x}$	$2.8 \cdot 10^{-4} \pm 1.2 \cdot 10^{-5}$	$3.4 \cdot 10^{-4} \pm 1.5 \cdot 10^{-5}$
std_{IC} on $\boldsymbol{v}$	$1.14 \cdot 10^{-3} \pm 7.8 \cdot 10^{-5}$	$2.7 \cdot 10^{-3} \pm 1.5 \cdot 10^{-4}$
std_{IC} on $\boldsymbol{y}$	$4.7 \cdot 10^{-6} \pm 3.0 \cdot 10^{-7}$	$6.2 \cdot 10^{-6} \pm 3.6 \cdot 10^{-7}$
mean_{IC} on $\boldsymbol{x}$	$2.22 \cdot 10^{-3} \pm 8.4 \cdot 10^{-5}$	$2.4 \cdot 10^{-3} \pm 1.0 \cdot 10^{-4}$
mean_{IC} on $\boldsymbol{v}$	$7.8 \cdot 10^{-3} \pm 3.8 \cdot 10^{-4}$	$1.78 \cdot 10^{-2} \pm 8.7 \cdot 10^{-4}$
mean_{IC} on $\boldsymbol{y}$	$1.91 \cdot 10^{-5} \pm 8.6 \cdot 10^{-7}$	$2.4 \cdot 10^{-5} \pm 1.1 \cdot 10^{-6}$
std_{IC} on $\boldsymbol{x}$	$3.0 \cdot 10^{-4} \pm 2.4 \cdot 10^{-5}$	$3.4 \cdot 10^{-4} \pm 1.3 \cdot 10^{-5}$
std_{IC} on $\boldsymbol{v}$	$1.15 \cdot 10^{-3} \pm 6.8 \cdot 10^{-5}$	$2.7 \cdot 10^{-3} \pm 1.5 \cdot 10^{-4}$
std_{IC} on $\boldsymbol{y}$	$4.7 \cdot 10^{-6} \pm 2.6 \cdot 10^{-7}$	$6.2 \cdot 10^{-6} \pm 3.1 \cdot 10^{-7}$

Table 2.6: $U(r) = \frac{r^{1.5}}{1.5}$: Trajectory Errors.

We maintain a 3-digit relative accuracy in estimating the position/velocity of the agents, even though for the interaction kernels, we are only able to maintain a 1-digit relative accuracy.

2.6 Conclusion and further directions

We have described a second-order model of interacting agents that incorporates multiple agent types, an environment, external forces, and multivariable interaction kernels. The inference procedure described exploits the structure of the system to achieve a learning rate that only depends on the dimension of the interaction kernels, which is much smaller than the full ambient dimension $(2d+1)N$. Our estimators are strongly consistent, and in fact have learning rates that are min-max optimal within the nonparametric class, under mild assumptions on the interaction kernels and the system. We described how one can relate the expected supremum error of the trajectories for the system driven by the estimated interaction kernels to the difference between the true interaction kernels and the estimated ones – this result gives strong

support to the use of our weighted L^2 norms as the correct way to measure performance and derive estimators. A detailed discussion of the full numerical algorithm, including the inverse problem derived from data and a coercivity condition to ensure learnability, along with complex examples, were presented and we showed how the formulation presented covers a very wide range of systems coming from many disciplines.

There are various ways that one could build on this work to handle different systems and for many of these further directions, the theoretical framework, techniques, and theorems presented here would be directly useful. In particular, one could consider second-order stochastic systems or a similar system but on a manifold, more complex environments, having more unknowns within the model beyond just the interaction kernels (say estimating the non-collective forces as well), identifying the best feature maps to model the data, and considering semiparametric problems where there are hidden parameters within the interaction kernels or other parts of the model that we wish to estimate along with the interaction kernels. The generality of the model and its broad coverage of models across the sciences, together with the scalability and performance of the algorithm, could inspire new models – both explicit equations and nonparametric estimators learned from data – which are theoretically justified and highly practical.

2.7 Control of trajectory error

Proof of Theorem 2.4.8. We introduce the function

$$F[\varphi^{EA}](\boldsymbol{x}, \dot{\boldsymbol{x}}, \boldsymbol{s}^E, \boldsymbol{s}^A) := \varphi^E(||\boldsymbol{x}||, \boldsymbol{s}^E)\boldsymbol{x} + \varphi^A(||\boldsymbol{x}||, \boldsymbol{s}^A)\dot{\boldsymbol{x}}$$

defined on $\mathbb{R}^{2d+p^E+p^A}$ for functions $\varphi^E \in L^\infty([0,R] \times \mathbb{S}^E), \varphi^A \in L^\infty([0,R] \times \mathbb{S}^A)$. Similarly, let $F[\varphi^\xi](\boldsymbol{x}, \xi, s^\xi) := \varphi^\xi(||\boldsymbol{x}||, \boldsymbol{s}^\xi)\xi$. By assumption, $\widehat{\boldsymbol{Y}}_0 = \boldsymbol{Y}_0$ and $\dot{\widehat{\boldsymbol{Y}}}_0 = \dot{\boldsymbol{Y}}_0$. For every $t \in [0,T]$, by the fundamental theorem of calculus and the triangle inequality,

we have

$$\begin{aligned}
\|\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t\|_{\mathcal{S}}^2 &= \sum_{j=1}^{k} \sum_{i \in C_j} \frac{1}{N_j} \Big\| \int_{u=0}^{t} \int_{s=0}^{u} (\ddot{\boldsymbol{x}}_i(s) - \ddot{\widehat{\boldsymbol{x}}}_i(s)) ds du \Big\|^2 \\
&\leq t^2 \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \int_{u=0}^{t} \int_{s=0}^{u} \|\ddot{\boldsymbol{x}}_i(s) - \ddot{\widehat{\boldsymbol{x}}}_i(s)\|^2 ds du \\
&= tp \int_{u=0}^{t} \int_{s=0}^{u} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) \\
&+ \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) + \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) \\
&- \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\widehat{\boldsymbol{X}}_s, \widehat{\boldsymbol{V}}_s, \widehat{\boldsymbol{\Xi}}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\widehat{\boldsymbol{X}}_s, \widehat{\boldsymbol{V}}_s, \widehat{\boldsymbol{\Xi}}_s)\|_{\mathcal{S}}^2 ds du \\
&\leq 2T^2 \int_{u=0}^{t} \int_{s=0}^{u} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds du \qquad (2.7.1) \\
&\quad + 2T^2 \int_{u=0}^{t} \int_{s=0}^{u} I ds du + \qquad (2.7.2) \\
&\quad 2T^2 \int_{u=0}^{t} \int_{s=0}^{u} \|\mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\widehat{\boldsymbol{X}}_s, \widehat{\boldsymbol{V}}_s, \widehat{\boldsymbol{\Xi}}_s)\|_{\mathcal{S}}^2 ds du.
\end{aligned}$$

First we introduce the convenient notations of

$$\boldsymbol{s}_{i\widehat{i'}}^{E} = \boldsymbol{s}_{(k_i,k_{i'})}^{E}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i, \widehat{\boldsymbol{x}}_{i'}, \dot{\widehat{\boldsymbol{x}}}_{i'}, \widehat{\xi}_{i'}), \qquad \widehat{\boldsymbol{s}_{ii'}^{E}} = \boldsymbol{s}_{(k_i,k_{i'})}^{E}(\widehat{\boldsymbol{x}}_i, \dot{\widehat{\boldsymbol{x}}}_i, \widehat{\xi}_i, \widehat{\boldsymbol{x}}_{i'}, \dot{\widehat{\boldsymbol{x}}}_{i'}, \widehat{\xi}_{i'}) \qquad (2.7.3)$$

with analogous formulae for $\boldsymbol{s}_{\widehat{i}i'}^{E}, \boldsymbol{s}_{\widehat{i}i'}^{A}, \boldsymbol{s}_{i\widehat{i'}}^{A}, \boldsymbol{s}_{i\widehat{i'}}^{\xi}, \boldsymbol{s}_{\widehat{i}i'}^{\xi}, \widehat{\boldsymbol{s}_{ii'}^{A}}, \widehat{\boldsymbol{s}_{ii'}^{\xi}}$.

Above we have introduced the term

$$I = \Big\| \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\widehat{\boldsymbol{X}}_s, \widehat{\boldsymbol{V}}_s, \widehat{\boldsymbol{\Xi}}_s) \Big\|_{\mathcal{S}}^2,$$

which can be expressed explicitly as

$$I = \Big\| \Big(\sum_{j'=1}^{K} \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\widehat{\phi}_{jj'}^{EA}](\boldsymbol{r}_{ii'}(s), \dot{\boldsymbol{r}}_{ii'}(s), \boldsymbol{s}_{ii'}^{E}(s), \boldsymbol{s}_{ii'}^{A}(s)) \qquad (2.7.4)$$

$$- F[\widehat{\phi}^{EA}_{jj'}](\widehat{\boldsymbol{r}}_{ii'}(s), \dot{\widehat{\boldsymbol{r}}}_{ii'}(s), \widehat{\boldsymbol{s}^E_{ii'}}(s), \widehat{\boldsymbol{s}^A_{ii'}}(s)))\Big)_{i,j}\Big\|^2_{\mathcal{S}}.$$

Note that in I, j is the index of the type among the $\{1, \dots K\}$ and i indexes within each type C_j. This holds similarly in later expressions I_1, I_2. For the third term of (2.7.1), we exploit the Lipschitz property of the non-collective force:

$$\begin{aligned}
&\|\mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\mathrm{nc},\dot{\widehat{\boldsymbol{x}}}}(\widehat{\boldsymbol{X}}_s, \widehat{\boldsymbol{V}}_s, \widehat{\boldsymbol{\Xi}}_s)\|^2_{\mathcal{S}} \\
&\qquad = \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \|\mathbf{F}_i^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i(s), \dot{\boldsymbol{x}}_i(s), \xi_i(s)) - \mathbf{F}_i^{\dot{\widehat{\boldsymbol{x}}}}(\widehat{\boldsymbol{x}}_i(s), \dot{\widehat{\boldsymbol{x}}}(s), \widehat{\xi}_i(s))\|^2 \\
&\qquad \leq \sum_{j=1}^{K} \sum_{i \in C_j} \frac{1}{N_j} \mathrm{Lip}^2[\mathbf{F}_i^{\dot{\boldsymbol{x}}}] \Big(\|\boldsymbol{x}_i(s) - \widehat{\boldsymbol{x}}_i(s)\|^2 \\
&\qquad + \|\dot{\boldsymbol{x}}_i(s) - \dot{\widehat{\boldsymbol{x}}}_i(s)\|^2 + \|\xi_i(s) - \widehat{\xi}_i(s)\|^2 \Big) \\
&\qquad \leq \max_i \mathrm{Lip}^2[\mathbf{F}_i^{\dot{\boldsymbol{x}}}] \|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|^2_{\mathcal{Y}}
\end{aligned}$$

So that we have the bound

$$2T^2 \int_{u=0}^{t} \int_{s=0}^{u} \|\mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\mathrm{nc},\dot{\widehat{\boldsymbol{x}}}}(\widehat{\boldsymbol{X}}_s, \widehat{\boldsymbol{V}}_s, \widehat{\boldsymbol{\Xi}}_s)\|^2_{\mathcal{S}} ds du \tag{2.7.5}$$

$$\leq 2T^2 \int_{u=0}^{t} \int_{s=0}^{u} \max_i \mathrm{Lip}^2[\mathbf{F}_i^{\dot{\boldsymbol{x}}}] \|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|^2_{\mathcal{Y}} ds du \tag{2.7.6}$$

Now we break up I using the triangle inequality and get that $I \leq I_1 + I_2$ where

$$\begin{aligned}
I_1 = \Bigg\| \Bigg(&\sum_{j'=1}^{K} \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\widehat{\phi}^{EA}_{jj'}](\boldsymbol{r}_{ii'}(s), \dot{\boldsymbol{r}}_{ii'}(s), \boldsymbol{s}^E_{ii'}(s), \boldsymbol{s}^A_{ii'}(s)) \\
&- F[\widehat{\phi}^{EA}_{jj'}](\boldsymbol{x}_i(s) - \widehat{\boldsymbol{x}_{i'}}(s), \dot{\boldsymbol{x}}_i(s) - \dot{\widehat{\boldsymbol{x}}}_i(s), \boldsymbol{s}^E_{i\widehat{i'}}(s), \boldsymbol{s}^A_{i\widehat{i'}}(s))) \Bigg)_{i,j} \Bigg\|^2_{\mathcal{S}}
\end{aligned}$$

$$I_2 = \Bigg\| \Bigg(\sum_{j'=1}^{K} \sum_{i' \in C_{j'}} \frac{1}{N_{j'}} (F[\widehat{\phi}^{EA}_{jj'}](\boldsymbol{x}_i(s) - \widehat{\boldsymbol{x}_{i'}}(s), \dot{\boldsymbol{x}}_i(s) - \dot{\widehat{\boldsymbol{x}}}_i(s), \boldsymbol{s}^E_{i\widehat{i'}}(s), \boldsymbol{s}^A_{i\widehat{i'}}(s))$$

$$- F_{[\widehat{\phi}^{EA}_{jj'}]}(\widehat{\boldsymbol{r}_{ii'}}(s), \widehat{\dot{\boldsymbol{r}}_{ii'}}(s), \widehat{\boldsymbol{s}^E_{ii'}}(s), \widehat{\boldsymbol{s}^A_{ii'}}(s)))\Big)_{i,j}\Big\|^2_{\mathcal{S}}$$

Using the Lipschitz property of $F[\widehat{\phi}^{EA}_{jj'}]$ we get that, since

$$I_1 = \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \Big| \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}} (F[\widehat{\phi}^{EA}_{jj'}](\boldsymbol{r}_{ii'}(s), \dot{\boldsymbol{r}}_{ii'}(s), \boldsymbol{s}^E_{ii'}(s), \boldsymbol{s}^A_{ii'}(s))$$
$$- F[\widehat{\phi}^{EA}_{jj'}](\boldsymbol{x}_i(s) - \widehat{\boldsymbol{x}_{i'}}(s), \dot{\boldsymbol{x}}_i(s) - \dot{\widehat{\boldsymbol{x}}}_i(s), \boldsymbol{s}^E_{i\widehat{i'}}(s), \boldsymbol{s}^A_{i\widehat{i'}}(s)))\Big|^2,$$

then,

$$I_1 \le K \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}} \Big| \Big(\mathrm{Lip}[F[\widehat{\phi}^{EA}_{jj'}]] \Big\| (\boldsymbol{x}_{i'}(s) - \widehat{\boldsymbol{x}_{i'}}(s), \dot{\boldsymbol{x}}_{i'}(s) - \dot{\widehat{\boldsymbol{x}}}_{i'}(s),$$
$$\boldsymbol{s}^E_{ii'}(s) - \boldsymbol{s}^E_{i\widehat{i'}}(s), \boldsymbol{s}^A_{ii'}(s) - \boldsymbol{s}^A_{i\widehat{i'}}(s)) \Big\| \Big) \Big|^2.$$

By the assumptions on the feature maps, we have that

$$\|\boldsymbol{s}^E_{ii'}(s) - \boldsymbol{s}^E_{i\widehat{i'}}(s)\| \le \mathrm{Lip}[\boldsymbol{s}^E_{(k_i,k_{i'})}] \|(\boldsymbol{x}_{i'}(s) - \widehat{\boldsymbol{x}_{i'}}(s), \dot{\boldsymbol{x}}_{i'}(s) - \dot{\widehat{\boldsymbol{x}}}_{i'}(s), \xi_{i'}(s) - \widehat{\xi}_{i'}(s))\|$$
$$\|\boldsymbol{s}^A_{ii'}(s) - \boldsymbol{s}^A_{i\widehat{i'}}(s)\| \le \mathrm{Lip}[\boldsymbol{s}^A_{(k_i,k_{i'})}] \|(\boldsymbol{x}_{i'}(s) - \widehat{\boldsymbol{x}_{i'}}(s), \dot{\boldsymbol{x}}_{i'}(s) - \dot{\widehat{\boldsymbol{x}}}_{i'}(s), \xi_{i'}(s) - \widehat{\xi}_{i'}(s))\|$$

Combining these bounds we see that,

$$I_1 \le K \sum_{j=1}^K \sum_{i\in C_j} \frac{1}{N_j} \sum_{j'=1}^K \sum_{i'\in C_{j'}} \frac{1}{N_{j'}}$$
$$\Big(\max_{j,j'} \Big(\mathrm{Lip}[F[\widehat{\phi}^{EA}_{jj'}](\mathrm{Lip}[\boldsymbol{s}^E_{(j,j')}] + 1), \mathrm{Lip}[F[\widehat{\phi}^{EA}_{jj'}](\mathrm{Lip}[\boldsymbol{s}^A_{(j,j')}] + 1) \Big) \Big)^2$$
$$(\|\boldsymbol{x}_{i'}(s) - \widehat{\boldsymbol{x}_{i'}}(s)\| + \|\dot{\boldsymbol{x}}_{i'}(s) - \dot{\widehat{\boldsymbol{x}}}_{i'}(s)\| + \|\xi_{i'}(s) - \widehat{\xi}_{i'}(s)\|)^2. \tag{2.7.7}$$

Let $\tilde{S} = \max(S_E, S_A)^2$, $J = (\max_{j,j'} \mathrm{Lip}[\boldsymbol{s}^E_{(j,j')}, \boldsymbol{s}^A_{(j,j')}] + 1)^2$, and then let $P = \tilde{S}J$ and

we get by Young's inequality that,

$$I_1 \leq 4KP\|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|_{\mathcal{Y}}^2, \tag{2.7.8}$$

and performing a similar analysis we get that

$$I_2 \leq 4KP\|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|_{\mathcal{Y}}^2.$$

So gathering terms, we can reexpress (2.7.1) as

$$\begin{aligned}\|\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t\|_{\mathcal{S}}^2 \leq 2T^2 \int_{u=0}^{t}\int_{s=0}^{u} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds du \\ + 2T^2(F + 8KP)\int_{u=0}^{t}\int_{s=0}^{u} \|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|_{\mathcal{Y}}^2 ds du\end{aligned} \tag{2.7.9}$$

where $F = \max_i \mathrm{Lip}[\mathbf{F}_i^{\dot{\boldsymbol{x}}}]$. Performing an analogous analysis on $\|\boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t\|_{\mathcal{S}}^2, \|\boldsymbol{\Xi}_t - \widehat{\boldsymbol{\Xi}}_t\|_{\mathcal{S}}^2$, with some additional effort, one can get the following result on the phase variable

$$\begin{aligned}\|\boldsymbol{\Xi}_t - \widehat{\boldsymbol{\Xi}}_t\|_{\mathcal{S}}^2 \leq 2T(8QK + F^{\xi})\int_{s=0}^{t} \|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|_{\mathcal{Y}}^2 ds \\ + 2T\int_{s=0}^{t} \|\dot{\boldsymbol{\Xi}}_s - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) + \mathbf{f}^{\phi^{\xi}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds\end{aligned} \tag{2.7.10}$$

where $F^{\xi} = \max_i \mathrm{Lip}[\mathbf{F}_i^{\xi}]$ and $Q = \max(H, S^{\xi})$ where $H = \max_{j,j'} \mathrm{Lip}[\boldsymbol{s}_{(j,j')}^{E}, \boldsymbol{s}_{(j,j')}^{A}]$. Similarly, we have that,

$$\begin{aligned}\|\boldsymbol{V}_t - \widehat{\boldsymbol{V}}_t\|_{\mathcal{S}}^2 \leq 2T\int_{s=0}^{t} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds \\ + 2T(F + 8KP)\int_{s=0}^{t} \|\boldsymbol{Y}_s - \widehat{\boldsymbol{Y}}_s\|_{\mathcal{Y}}^2 ds\end{aligned} \tag{2.7.11}$$

Gathering the bounds (2.7.9, 2.7.10, 2.7.11), we have that

$$\begin{aligned}\|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2 &\leq 2T(8KP + F + 8QK + F^{\xi}) \int_{s=0}^{t} \|\widehat{\boldsymbol{Y}}_s - \boldsymbol{Y}_s\|_{\mathcal{Y}}^2 ds \\ &\quad + 2T^2(8KP + F) \int_{u=0}^{t} \int_{s=0}^{u} \|\widehat{\boldsymbol{Y}}_s - \boldsymbol{Y}_s\|_{\mathcal{Y}}^2 ds du \\ &\quad + a(t) \left\{ \begin{array}{l} +2T^2 \displaystyle\int_{u=0}^{t} \int_{s=0}^{u} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds du \\ +2T \displaystyle\int_{s=0}^{t} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds \\ +2T \displaystyle\int_{s=0}^{t} \|\dot{\boldsymbol{\Xi}} - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\phi}^{\xi}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds \end{array} \right] \end{aligned}$$

where we denote the last three lines by $a(t)$ and notice that this is a nondecreasing function in t. We also denote $A_1 = 2T(8KP+F+8QK+F^{\xi})$ and $B_1 = 2T^2(8KP+F)$. Now use theorem 2.11.1, which is in [34] and is originally in Bainov and Simeonov. With this notation, we can rewrite the above bound as

$$\|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2 \leq A_1 \int_{s=0}^{t} \|\widehat{\boldsymbol{Y}}_s - \boldsymbol{Y}_s\|_{\mathcal{Y}}^2 ds + B_1 \int_{u=0}^{t} \int_{s=0}^{u} \|\widehat{\boldsymbol{Y}}_s - \boldsymbol{Y}_s\|_{\mathcal{Y}}^2 ds du + a(t) \quad (2.7.12)$$

And so in the notation of Theorem 2.11.1 we have $u(t) = \|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2$, $b(t) = 1$, $k_1(t, t_1) = A_1$ and $k_2(t, t_1, t_2) = B_1$, so that for all t we have

$$\|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2 \leq a(t) + \int_0^t \widehat{R}[a](t, s) \exp\left(\int_s^t \widehat{R}[b](t, \tau) d\tau \right) ds$$

and we have the simple bounds

$$\widehat{R}[a](t, s) = a(t) + \int_0^s B_1 a(t_2) dt_2 \leq a(T) + B_1 T a(T) \quad (2.7.13)$$

$$\widehat{R}[b](t, \tau) = A_1 + \int_0^{\tau} 1 dy = A_1 + \tau \quad (2.7.14)$$

So that,

$$\begin{aligned}\|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2 &\leq a(T) + [a(T) + B_1 T a(T)] \int_{s=0}^{t} \exp\left(\int_s^t (A_1 + \tau) d\tau\right) ds \\ &\leq a(T) + [a(T) + B_1 T a(T)] \int_{s=0}^{T} \exp\left(\int_0^T A_1 + \tau d\tau\right) ds \\ &= a(T) + [a(T) + B_1 T a(T)] T \exp(A_1 T + T^2/2) \\ &= a(T)(1 + (1 + B_1 T) T \exp(A_1 T + T^2/2))\end{aligned}$$

So that we can immediately conclude the first assertion of the theorem,

$$\sup_{t\in[0,T]} \|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Y}_t\|_{\mathcal{Y}}^2 \leq a(T)(1 + (T + B_1 T^2) \exp(A_1 T + T^2/2))$$

Lastly, we can use the results of section 2.8.1 to get the key result on the expected supremum error. We take expectation on each of the three terms of $a(T)$ and normalize them so they are in the form of the results of 2.8.1.

$$\begin{aligned}\frac{1}{T^2} \int_{u=0}^{T} \int_{s=0}^{T} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \|\ddot{\boldsymbol{X}}_s - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\boldsymbol{\phi}}^{EA}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds du \\ \leq K^2 \|\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA})}^2 .\end{aligned}$$

We similarly get that,

$$\frac{1}{T} \int_{s=0}^{T} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \|\dot{\boldsymbol{\Xi}}_s - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s) - \mathbf{f}^{\widehat{\boldsymbol{\phi}}^{\xi}}(\boldsymbol{X}_s, \boldsymbol{V}_s, \boldsymbol{\Xi}_s)\|_{\mathcal{S}}^2 ds \leq K^2 \|\widehat{\boldsymbol{\phi}}^{\xi} - \boldsymbol{\phi}^{\xi}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi})}^2 , \tag{2.7.15}$$

and can get an analogous bound for the remaining term of $a(T)$. These bounds together lead to

$$a(T) \leq (2T^4 K^2 + 2T^2 K^2) \|\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA})}^2 + 2T^2 K^2 \|\widehat{\boldsymbol{\phi}}^{\xi} - \boldsymbol{\phi}^{\xi}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi})}^2$$

which implies the desired result. □

2.8 Learning theory - technical tools

2.8.1 Continuity of the error functionals

For any $t \in [0,T]$, consider the two random variables,

$$\begin{aligned}\boldsymbol{\mathcal{E}}_{\boldsymbol{X}_t}^{EA}(\boldsymbol{\varphi}^{EA}) = \Big\| \ddot{\boldsymbol{X}}_t - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) \\ - \mathbf{f}^{\boldsymbol{\varphi}^E}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) - \mathbf{f}^{\boldsymbol{\varphi}^A}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) \Big\|_{\mathcal{S}}^2 \end{aligned} \tag{2.8.1}$$

$$\boldsymbol{\mathcal{E}}_{\boldsymbol{\Xi}_t}^{\xi}(\boldsymbol{\varphi}^{\xi}) = \Big\| \dot{\boldsymbol{\Xi}}_t - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) - \mathbf{f}^{\boldsymbol{\varphi}^{\xi}}(\boldsymbol{X}_t, \boldsymbol{V}_t, \boldsymbol{\Xi}_t) \Big\|_{\mathcal{S}}^2 \tag{2.8.2}$$

These will be used in various places throughout the technical proofs and easily relate to the error functionals

$$\begin{aligned}\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA}) &:= \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \frac{1}{L} \sum_{l=1}^{L} \Big\| \ddot{\boldsymbol{X}}_{t_l} - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) - \mathbf{f}^{\boldsymbol{\varphi}^E}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) - \\ &\qquad \mathbf{f}^{\boldsymbol{\varphi}^A}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) \Big\|_{\mathcal{S}}^2 \\ \boldsymbol{\mathcal{E}}_{\infty}^{\xi}(\boldsymbol{\varphi}^{\xi}) &:= \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \frac{1}{L} \sum_{l=1}^{L} \Big[\Big\| \dot{\boldsymbol{\Xi}}_{t_l} - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) - \mathbf{f}^{\boldsymbol{\varphi}^{\xi}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) \Big\|_{\mathcal{S}}^2 \Big], \end{aligned} \tag{2.8.3}$$

which by the Strong Law of Large Numbers satisfy

$$\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA}) = \lim_{M \to \infty} \boldsymbol{\mathcal{E}}_{M}^{EA}(\boldsymbol{\varphi}^{EA}) \quad \text{and} \quad \boldsymbol{\mathcal{E}}_{\infty}^{\xi}(\boldsymbol{\varphi}^{\xi}) = \lim_{M \to \infty} \boldsymbol{\mathcal{E}}_{M}^{\xi}(\boldsymbol{\varphi}^{\xi}) .$$

Indeed

$$\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA}) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left[\boldsymbol{\mathcal{E}}_{\boldsymbol{X}_{t_l}}^{EA}(\boldsymbol{\varphi}^{EA}) \right], \qquad \boldsymbol{\mathcal{E}}_{\infty}^{\xi}(\boldsymbol{\varphi}^{\xi}) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left[\boldsymbol{\mathcal{E}}_{\boldsymbol{\Xi}_{t_l}}^{\xi}(\boldsymbol{\varphi}^{\xi}) \right].$$

We begin by establishing basic continuity results for our error functionals over the hypothesis space. The specific structure of the governing equations plays a critical role in the analysis.

Alignment and energy based kernels

Proposition 2.8.1. *For $\widehat{\boldsymbol{\varphi}}^{EA}, \widehat{\boldsymbol{\phi}}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$ the true and empirical error functionals are bounded as follows,*

$$|\boldsymbol{\mathcal{E}}^{EA}_{\infty}(\widehat{\boldsymbol{\varphi}}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_{\infty}(\widehat{\boldsymbol{\phi}}^{EA})| \leq K^2 \|\widehat{\boldsymbol{\varphi}}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \|2\boldsymbol{\phi}^{EA} - \widehat{\boldsymbol{\varphi}}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \tag{2.8.4}$$

$$|\boldsymbol{\mathcal{E}}^{EA}_{M}(\widehat{\boldsymbol{\varphi}}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_{M}(\widehat{\boldsymbol{\phi}}^{EA})| \leq K^4 \max\{R, R_{\dot{x}}\}^2 \|\widehat{\boldsymbol{\varphi}}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}\|_{\infty} \|2\boldsymbol{\phi}^{EA} - \widehat{\boldsymbol{\varphi}}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}\|_{\infty} \tag{2.8.5}$$

Recall the definitions of $R, R_{\dot{x}}$ in equations (2.3.7)*, and* (2.3.9)*.*

Proof. Using Jensen's inequality,

$$\begin{aligned}
&|\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}_t}(\widehat{\boldsymbol{\varphi}}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}_t}(\widehat{\boldsymbol{\phi}}^{EA})| \\
&= \Bigg| \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i \in C_k} \Big\langle \sum_{k'=1}^{K} \frac{1}{N_{k'}} \sum_{i' \in C_{k'}} (\widehat{\varphi}^E_{kk'} - \widehat{\phi}^E_{kk'})(r_{ii'}, \boldsymbol{s}^E_{ii'}) \boldsymbol{r}_{ii'} + (\widehat{\varphi}^A_{kk'} - \widehat{\phi}^A_{kk'})(r_{ii'}, \boldsymbol{s}^A_{ii'}) \dot{\boldsymbol{r}}_{ii'}, \\
&\sum_{k''=1}^{K} \frac{1}{N_{k''}} \sum_{i' \in C_{k''}} (2\phi^E_{kk'} - \widehat{\varphi}^E_{kk'} - \widehat{\phi}^E_{kk'})(r_{ii'}, \boldsymbol{s}^E_{ii'}) \boldsymbol{r}_{ii'} + (2\phi^A_{kk'} - \widehat{\varphi}^A_{kk'} - \widehat{\phi}^A_{kk'})(r_{ii'}, \boldsymbol{s}^A_{ii'}) \dot{\boldsymbol{r}}_{ii'} \Big\rangle \Bigg| \\
&\leq \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{k''=1}^{K} \frac{1}{N_k} \sum_{i \in C_k} \Big\| \frac{1}{N_{k'}} \sum_{i' \in C_{k'}} (\widehat{\varphi}^E_{kk'} - \widehat{\phi}^E_{kk'})(r_{ii'}, \boldsymbol{s}^E_{ii'}) \boldsymbol{r}_{ii'} + (\widehat{\varphi}^A_{kk'} - \widehat{\phi}^A_{kk'})(r_{ii'}, \boldsymbol{s}^A_{ii'}) \dot{\boldsymbol{r}}_{ii'} \Big\|
\end{aligned} \tag{2.8.6}$$

$$\Big\| \frac{1}{N_{k''}} \sum_{i' \in C_{k''}} (2\phi^E_{kk'} - \widehat{\varphi}^E_{kk'} - \widehat{\phi}^E_{kk'})(r_{ii'}, \boldsymbol{s}^E_{ii'}) \boldsymbol{r}_{ii'} + (2\phi^A_{kk'} - \widehat{\varphi}^A_{kk'} - \widehat{\phi}^A_{kk'})(r_{ii'}, \boldsymbol{s}^A_{ii'}) \dot{\boldsymbol{r}}_{ii'} \Big\|$$

$$\leq \sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{k''=1}^{K}\sqrt{\frac{1}{N_kN_{k'}}\sum_{i\in C_k,i'\in C_{k'}}\|(\widehat{\varphi}^E_{kk'}-\widehat{\phi}^E_{kk'})(r_{ii'},\boldsymbol{s}^E_{ii'})\boldsymbol{r}_{ii'}+(\widehat{\varphi}^A_{kk'}-\widehat{\phi}^A_{kk'})(r_{ii'},\boldsymbol{s}^A_{ii'})\dot{\boldsymbol{r}}_{ii'}\|^2}\times$$
$$\sqrt{\frac{1}{N_kN_{k''}}\sum_{i\in C_k,i'\in C_{k''}}\|(2\phi^E_{kk'}-\widehat{\varphi}^E_{kk'}-\widehat{\phi}^E_{kk'})(r_{ii'},\boldsymbol{s}^E_{ii'})\boldsymbol{r}_{ii'}+(2\phi^A_{kk'}-\widehat{\varphi}^A_{kk'}-\widehat{\phi}^A_{kk'})(r_{ii'},\boldsymbol{s}^A_{ii'})\dot{\boldsymbol{r}}_{ii'}\|^2}$$
$$\leq \sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{k''=1}^{K}\|(\widehat{\varphi}^{EA}_{kk'}-\widehat{\phi}^{EA}_{kk'})\|_{L^2(\hat{\rho}_T^{t,kk'})}\|2(\phi^{EA}_{kk'}-\widehat{\varphi}^{EA}_{kk'}-\widehat{\varphi}^{EA}_{kk'})\|_{L^2(\hat{\rho}_T^{t,kk''})}$$
$$\leq K^2\|\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}^t_T)}\|2\boldsymbol{\phi}^{EA}-\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}^t_T)}, \tag{2.8.7}$$

where

$$\widehat{\rho}_T^{t,kk'}(r,\boldsymbol{s}^E,\dot{r},\boldsymbol{s}^A)=\frac{1}{LN_{kk'}}\sum_{l=1}^{L}\sum_{\substack{i\in C_k,i'\in C_{k'}\\ i\neq i'}}\delta_{r_{ii'}(t),\boldsymbol{s}^E_{ii'}(t_l),\dot{r}_{ii'}(t_l),\boldsymbol{s}^A_{ii'}(t_l)}(r,\boldsymbol{s}^E,\dot{r},\boldsymbol{s}^A)$$

and $\widehat{\boldsymbol{\rho}}^t_T=\bigoplus_{k,k'=1,1}^{K,K}\widehat{\rho}_T^{t,kk'}$. Therefore, we have that

$$\big|\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\varphi}}^{EA})-\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}^{EA})\big|\leq\frac{1}{L}\sum_{l=1}^{L}\big|\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}^{EA})-\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\varphi}}^{EA})\big|$$
$$<\frac{K^2}{L}\sum_{l=1}^{L}\|\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}^{t_l}_T)}\|2\boldsymbol{\phi}^{EA}-\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}^{t_l}_T)}$$
$$\leq K^2\sqrt{\frac{1}{L}\sum_{l=1}^{L}\|\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|^2_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}^{t_l}_T)}}\sqrt{\frac{1}{L}\sum_{l=1}^{L}\|2\boldsymbol{\phi}^{EA}-\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|^2_{\boldsymbol{L}^2(\hat{\boldsymbol{\rho}}^{t_l}_T)}}$$
$$=K^2\|\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}^L_T)}\|2\boldsymbol{\phi}^{EA}-\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\widehat{\boldsymbol{\rho}}^L_T)} \tag{2.8.8}$$
$$\leq K^4\max\{R,R_{\dot{x}}\}^2\|\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_\infty\|2\boldsymbol{\phi}^{EA}-\widehat{\boldsymbol{\varphi}}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}\|_\infty \tag{2.8.9}$$

Taking the expectation with respect to $\boldsymbol{\mu}^{\boldsymbol{Y}}$ on each side of (2.8.8) we get the first inequality. The second inequality follows by noticing that,

$$|\boldsymbol{\mathcal{E}}^{EA}_M(\widehat{\boldsymbol{\varphi}}^{EA})-\boldsymbol{\mathcal{E}}^{EA}_M(\widehat{\boldsymbol{\phi}}^{EA})|\leq\frac{1}{M}\sum_{m=1}^{M}\big|\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mathcal{E}}_{\boldsymbol{X}^{(m)}_{t_l}}(\widehat{\boldsymbol{\varphi}}^{EA})-\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\mathcal{E}}_{\boldsymbol{X}^{(m)}_{t_l}}(\widehat{\boldsymbol{\phi}}^{EA})\big|.$$

□

Environment interaction kernels

Here we show an analogous result to the alignment and energy result above. The techniques are similar and the result serves an identical purpose in the theory. Recall the definition of R_ξ in (2.3.10).

Proposition 2.8.2. *For $\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\phi}} \in \boldsymbol{\mathcal{H}}^\xi$, we have*

$$|\boldsymbol{\mathcal{E}}^\xi_\infty(\widehat{\boldsymbol{\varphi}}) - \boldsymbol{\mathcal{E}}^\xi_\infty(\widehat{\boldsymbol{\phi}})| \leq K^2 \|\widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\phi}}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}^{\xi,L}_T)} \|2\boldsymbol{\phi}^\xi - \widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\phi}}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}^{\xi,L}_T)} \tag{2.8.10}$$

$$|\boldsymbol{\mathcal{E}}^\xi_M(\widehat{\boldsymbol{\varphi}}) - \boldsymbol{\mathcal{E}}^\xi_M(\widehat{\boldsymbol{\phi}})| \leq K^4 R^2_\xi \|\widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\phi}}\|_\infty \|2\boldsymbol{\phi} - \widehat{\boldsymbol{\varphi}} - \widehat{\boldsymbol{\phi}}\|_\infty \tag{2.8.11}$$

The following lemma can be immediately deduced using (2.8.4), (2.8.5) , and (2.8.9) .

Lemma 2.8.3. *For all $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$, define the defect function $L^{EA}_M(\boldsymbol{\varphi}^{EA})$ as*

$$L^{EA}_M(\boldsymbol{\varphi}^{EA}) = \boldsymbol{\mathcal{E}}^{EA}_\infty(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_M(\boldsymbol{\varphi}^{EA}). \tag{2.8.12}$$

Then, given two functions $\boldsymbol{\varphi}^{EA}_1, \boldsymbol{\varphi}^{EA}_2 \in \boldsymbol{\mathcal{H}}^{EA}$, the defect function is bounded by

$$|L^{EA}_M(\boldsymbol{\varphi}^{EA}_1) - L^{EA}_M(\boldsymbol{\varphi}^{EA}_2)| \leq 2K^4 \max\{R, R_{\dot{x}}\}^2 \|\boldsymbol{\varphi}^{EA}_1 - \boldsymbol{\varphi}^{EA}_2\|_\infty \|\boldsymbol{\varphi}^{EA}_1 + \boldsymbol{\varphi}^{EA}_2 - 2\boldsymbol{\phi}^{EA}\|_\infty$$

almost surely with respect to $\boldsymbol{\mu}^{\boldsymbol{Y}}$.

A similar lemma can be immediately deduced on the ξ variable .

Lemma 2.8.4. *For all $\boldsymbol{\varphi}^\xi \in \boldsymbol{\mathcal{H}}^\xi$, define the defect function $L^\xi_M(\boldsymbol{\varphi}^\xi)$ as*

$$L^\xi_M(\boldsymbol{\varphi}^\xi) = \boldsymbol{\mathcal{E}}^\xi_\infty(\boldsymbol{\varphi}^\xi) - \boldsymbol{\mathcal{E}}^\xi_M(\boldsymbol{\varphi}^\xi). \tag{2.8.13}$$

Then, given two functions $\boldsymbol{\varphi}_1^{\xi}, \boldsymbol{\varphi}_2^{\xi} \in \boldsymbol{\mathcal{H}}^{\xi}$, the defect function is bounded by

$$|L_M^{\xi}(\boldsymbol{\varphi}_1^{\xi}) - L_M^{\xi}(\boldsymbol{\varphi}_2^{\xi})| \leq 2K^4 R_{\xi}^2 \|\boldsymbol{\varphi}_1^{\xi} - \boldsymbol{\varphi}_2^{\xi}\|_{\infty} \|\boldsymbol{\varphi}_1^{\xi} + \boldsymbol{\varphi}_2^{\xi} - 2\boldsymbol{\phi}^{\xi}\|_{\infty}$$

almost surely with respect to $\boldsymbol{\mu}^{\boldsymbol{Y}}$.

2.8.2 Uniqueness of minimizers over a compact convex space

Recall the energy and alignment bilinear functional $\langle\langle \cdot, \cdot \rangle\rangle_{EA}$, previously defined in equation (2.4.7)

$$\langle\langle \boldsymbol{\varphi}_1^{EA}, \boldsymbol{\varphi}_2^{EA} \rangle\rangle_{EA} := \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left[\left\langle \mathbf{f}_{\boldsymbol{\varphi}_1^{EA}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}), \mathbf{f}_{\boldsymbol{\varphi}_2^{EA}}(\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \boldsymbol{\Xi}_{t_l}) \right\rangle_{\mathcal{S}} \right],$$

for any $\boldsymbol{\varphi}_1^{EA}, \boldsymbol{\varphi}_2^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$. The $\mathcal{S}$-inner product is the inner product induced by the $\|\cdot\|_{\mathcal{S}}$ norm by the polarization identity, which holds as we are working in an L^2 space, so the parallelogram law holds. Then our coercivity condition (2.4.5). can be written in terms of this bilinear functional as: for all $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$

$$c_{\boldsymbol{\mathcal{H}}^{EA}} \left\| \boldsymbol{\varphi}^{EA} \right\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}^2 \leq \langle\langle \boldsymbol{\varphi}^{EA}, \boldsymbol{\varphi}^{EA} \rangle\rangle$$

Proposition 2.8.5. *Let the minimizer of the error functional be denoted*

$$\widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}^{EA} := \widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}^{E} \oplus \widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}^{A} := \underset{\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}}{\arg\min}\ \boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA});$$

then for all $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$, the difference of the error functional at this element of $\boldsymbol{\mathcal{H}}^{EA}$ and the minimizer is lower bounded as,

$$\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}_{\infty}^{EA}(\widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}^{EA}) \geq c_{\boldsymbol{\mathcal{H}}^{EA}} \|\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})}^2 . \tag{2.8.14}$$

Thus, the minimizer of $\boldsymbol{\mathcal{E}}_{\infty}^{EA}$ *over* $\boldsymbol{\mathcal{H}}^{EA}$ *is unique in* $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})$.

Proof. For $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$, and to ease the notation let $\widehat{\boldsymbol{\phi}}^{EA} := \widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}$, we have

$$\begin{aligned}\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}_{\infty}^{EA}(\widehat{\boldsymbol{\phi}}^{EA}) &= \langle\!\langle(\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA}), (\boldsymbol{\varphi}^{EA} - \boldsymbol{\phi}^{EA})\rangle\!\rangle \\ &\quad - \langle\!\langle(\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}), (\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA})\rangle\!\rangle \qquad (2.8.15)\end{aligned}$$

Using that $\langle\!\langle X, X\rangle\!\rangle - \langle\!\langle Y, Y\rangle\!\rangle = \langle\!\langle X - Y, X + Y\rangle\!\rangle$, which holds by bilinearity and the definition of the form, we get that

$$\begin{aligned}(2.8.15) &= \langle\!\langle(\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}), (\boldsymbol{\varphi}^{EA} + \widehat{\boldsymbol{\phi}}^{EA} - 2\boldsymbol{\phi}^{EA}\rangle\!\rangle \\ &= \langle\!\langle(\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}), (\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA} + 2(\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA}))\rangle\!\rangle \\ &= \langle\!\langle(\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}), (\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA})\rangle\!\rangle + 2\langle\!\langle(\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}), (\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA})\rangle\!\rangle \\ &\qquad (2.8.16)\end{aligned}$$

By the coercivity condition, the first term in (2.8.16) is at least as large as

$$c_{\boldsymbol{\mathcal{H}}^{EA}} \|\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}\|_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{EA,L})} \geq 0$$

We are left to show the second term in (2.8.16) is nonnegative. Since $\boldsymbol{\mathcal{H}}^{EA}$ is convex, for all $t \in [0,1]$, $t(\boldsymbol{\varphi}^{EA}) + (1-t)(\widehat{\boldsymbol{\phi}}^{EA}) \in \boldsymbol{\mathcal{H}}^{EA}$. By definition of $\widehat{\boldsymbol{\phi}}^{EA}$ as an argmin,

$$\boldsymbol{\mathcal{E}}_{\infty}^{EA}(t\boldsymbol{\varphi}^{EA} + (1-t)\widehat{\boldsymbol{\phi}}^{EA}) - \boldsymbol{\mathcal{E}}_{\infty}^{EA}(\widehat{\boldsymbol{\phi}}^{EA}) \geq 0$$

which means, using a decomposition analogous to the one above in (2.8.16), that

$$\begin{aligned}&\langle\!\langle\!\langle(t\boldsymbol{\varphi}^{EA} + (1-t)\widehat{\boldsymbol{\phi}}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}), (t\boldsymbol{\varphi}^{EA} + (1-t)\widehat{\boldsymbol{\phi}}^{EA} - \widehat{\boldsymbol{\phi}}^{EA} + 2(\widehat{\boldsymbol{\phi}}^{EA} - \boldsymbol{\phi}^{EA})\rangle\!\rangle\!\rangle \\ &\qquad = \langle\!\langle\!\langle t((\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA})), (t\boldsymbol{\varphi}^{EA} + (2-t)\widehat{\boldsymbol{\phi}}^{EA} - 2\boldsymbol{\phi}^{EA})\rangle\!\rangle\!\rangle \geq 0\,.\end{aligned}$$

Therefore,

$$t\langle\!\langle(\boldsymbol{\varphi}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}),(t\boldsymbol{\varphi}^{EA}+(2-t)\widehat{\boldsymbol{\phi}}^{EA}-2\boldsymbol{\phi}^{EA})\rangle\!\rangle \geq 0 \tag{2.8.17}$$

$$\Leftrightarrow \langle\!\langle(\boldsymbol{\varphi}^{EA}-\widehat{\boldsymbol{\phi}}^{EA}),(t\boldsymbol{\varphi}^{EA}+(2-t)\widehat{\boldsymbol{\phi}}^{EA}-2\boldsymbol{\phi}^{EA})\rangle\!\rangle \geq 0 \tag{2.8.18}$$

By the results of section 2.8.1, we have (Lipschitz) continuity of the bilinear functional $\langle\!\langle\cdot,\cdot\rangle\!\rangle$ over $\boldsymbol{\mathcal{H}}^{EA}\times\boldsymbol{\mathcal{H}}^{EA}$. Next, take the $\lim_{t\to 0^+}$ of (2.8.17) and by the dominated convergence theorem (which holds due to the boundedness and continuity assumptions on the interaction kernels) we pass the limit through the expectations in $\langle\!\langle\cdot,\cdot\rangle\!\rangle$. This gives that (2.8.16) is greater than 0, giving the desired result on the uniqueness of the minimizer. □

Proposition 2.8.6. *Let the minimizer of the error functional be denoted*

$$\widehat{\boldsymbol{\phi}}_{L,\infty,\boldsymbol{\mathcal{H}}^{\xi}} := \underset{\boldsymbol{\varphi}^{\xi}\in\boldsymbol{\mathcal{H}}^{\xi}}{\arg\min}\,\boldsymbol{\mathcal{E}}_{\infty}^{\xi}(\boldsymbol{\varphi}^{\xi});$$

then for all $\boldsymbol{\varphi}^{\xi}\in\boldsymbol{\mathcal{H}}^{\xi}$, the difference of the error functional at this element of $\boldsymbol{\mathcal{H}}^{\xi}$ and the minimizer is lower bounded as,

$$\boldsymbol{\mathcal{E}}_{\infty}^{\xi}(\boldsymbol{\varphi}^{\xi})-\boldsymbol{\mathcal{E}}_{\infty}^{\xi}(\widehat{\boldsymbol{\phi}}^{\xi}_{L,\infty,\boldsymbol{\mathcal{H}}^{\xi}}) \geq c_{\boldsymbol{\mathcal{H}}^{\xi}}\|\boldsymbol{\varphi}^{\xi}-\widehat{\boldsymbol{\phi}}^{\xi}_{L,\infty,\boldsymbol{\mathcal{H}}^{\xi}}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi,L})}\,. \tag{2.8.19}$$

Thus, the minimizer of $\boldsymbol{\mathcal{E}}_{\infty}^{\xi}$ over $\boldsymbol{\mathcal{H}}^{\xi}$ is unique in $\boldsymbol{L}^2(\boldsymbol{\rho}_T^{\xi,L})$.

2.8.3 Uniform estimates on defect functions

We start this section by introducing normalized errors of the estimators. Denote the minimizer of $\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\cdot)$ over $\boldsymbol{\mathcal{H}}^{EA}$ by

$$\widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}} := \widehat{\boldsymbol{\phi}}^{E}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}\oplus\widehat{\boldsymbol{\phi}}^{A}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}} = \underset{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}}{\arg\min}\,\boldsymbol{\mathcal{E}}_{\infty}^{EA}(\boldsymbol{\varphi}^{EA}). \tag{2.8.20}$$

For any $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$, define the *normalized errors* as

$$\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) := \boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}})\,, \tag{2.8.21}$$

$$\mathcal{D}_M(\boldsymbol{\varphi}^{EA}) := \boldsymbol{\mathcal{E}}_M^{EA}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}_M^{EA}(\widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}})\,. \tag{2.8.22}$$

These quantities capture the difference between the expected/empirical errors of the estimator and the function in the hypothesis space minimizing the expected error functional. We begin by proving a lemma that assumes the distance between the expected and empirical normalized errors are small for a given estimator. We then show that we have similar control on these distances for all points in a neighborhood of this particular one. This control enables us to apply a covering argument in the main proposition of this section due to the compactness of the hypothesis space.

Remark 2.8.7. *Exactly analogous definitions hold for the ξ variable and we will simply state the results in that case.*

Lemma 2.8.8. *For all $\epsilon > 0$ and $0 < \alpha < 1$, if the function $\boldsymbol{\varphi}_1^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$ satisfies*

$$\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}_1^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) + \epsilon} < \alpha\,,$$

then for all $\boldsymbol{\varphi}_2^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$ such that $\|\boldsymbol{\varphi}_1^{EA} - \boldsymbol{\varphi}_2^{EA}\|_\infty \leq \frac{\alpha\epsilon}{8S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}$, where $S_{EA} = \max\{S_E, S_A\}$ we have

$$\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}_2^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} < 3\alpha.$$

Proof. To ease the notation, write $\widehat{\boldsymbol{\phi}}^{EA} := \widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}$, and using definition (2.8.12), we have that

$$\begin{aligned}\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}_2^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} &= \frac{\boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}_2^{EA}) - \boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}^{EA}) - (\boldsymbol{\mathcal{E}}_M^{EA}(\boldsymbol{\varphi}_2^{EA}) - \boldsymbol{\mathcal{E}}_M^{EA}(\widehat{\boldsymbol{\phi}}^{EA}))}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} \\ &= \frac{L_M^{EA}(\boldsymbol{\varphi}_2^{EA}) - L_M^{EA}(\boldsymbol{\varphi}_1^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} + \frac{L_M^{EA}(\boldsymbol{\varphi}_1^{EA}) - L_M^{EA}(\widehat{\boldsymbol{\phi}}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon}\end{aligned}$$

By Lemma 2.8.3, we have

$$L_M^{EA}(\boldsymbol{\varphi}_2^{EA}) - L_M^{EA}(\boldsymbol{\varphi}_1^{EA}) \leq 8S_{EA}\max\{R, R_{\dot{x}}\}^2 K^4 \|\boldsymbol{\varphi}_2^{EA} - \boldsymbol{\varphi}_1^{EA}\|_\infty \leq \alpha\epsilon.$$

By definition we have that $\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) \geq 0$ implying that,

$$\frac{L_M(\boldsymbol{\varphi}_1^{EA}) - L_M(\boldsymbol{\varphi}_2^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} \leq \alpha.$$

For the second term, by equation (2.8.5) and the assumption that $\alpha < 1$, we obtain that

$$\boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}_1^{EA}) - \boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}_2^{EA}) < 4S_{EA}\max\{R, R_{\dot{x}}\}^2 K^4 \|\boldsymbol{\varphi}_1^{EA} - \boldsymbol{\varphi}_2^{EA}\|_\infty < \epsilon.$$

Therefore

$$\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) - \mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) = \boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}_1^{EA}) - \boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}_2^{EA}) < \epsilon \leq \epsilon + \mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}),$$

and thus

$$\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) + \epsilon}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} \leq 2.$$

We conclude that

$$\frac{L_M^{EA}(\boldsymbol{\varphi}_1^{EA}) - L_M^{EA}(\widehat{\boldsymbol{\phi}}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} = \frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}_1^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) + \epsilon} \frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_1^{EA}) + \epsilon}{\mathcal{D}_\infty(\boldsymbol{\varphi}_2^{EA}) + \epsilon} < 2\alpha,$$

and the result follows by summing the two estimates. □

Arguing in the same way as above, we can derive the lemma below using equation 2.8.11. We define $\mathcal{D}_\infty^\xi, \mathcal{D}_M^\xi$ similarly to (2.8.21), (2.8.22) using $\boldsymbol{\mathcal{E}}_\infty^\xi, \boldsymbol{\mathcal{E}}_M^\xi$ in the obvious way.

Lemma 2.8.9. *For all $\epsilon > 0$ and $0 < \alpha < 1$, if the function $\boldsymbol{\phi}_1^\xi \in \boldsymbol{\mathcal{H}}^\xi$ satisfies*

$$\frac{\mathcal{D}_\infty^\xi(\boldsymbol{\phi}_1^\xi) - \mathcal{D}_M^\xi(\boldsymbol{\phi}_1^\xi)}{\mathcal{D}_\infty^\xi(\boldsymbol{\phi}_1^\xi) + \epsilon} < \alpha\,,$$

then for all $\boldsymbol{\phi}_2^\xi \in \boldsymbol{\mathcal{H}}^\xi$ such that $\|\boldsymbol{\phi}_1^\xi - \boldsymbol{\phi}_2^\xi\|_\infty \leq \frac{\alpha\epsilon}{8S_0R_\xi^2K^4}$, we have, for $S_0 \geq S_\xi$,

$$\frac{\mathcal{D}_\infty^\xi(\boldsymbol{\phi}_2^\xi) - \mathcal{D}_M^\xi(\boldsymbol{\phi}_2^\xi)}{\mathcal{D}_\infty^\xi(\boldsymbol{\phi}_2^\xi) + \epsilon} < 3\alpha.$$

2.8.4 Concentration

Proposition 2.8.10. *For all $\epsilon > 0$, $0 < \alpha < 1$, $\boldsymbol{\varphi}^{EA} \in \boldsymbol{\mathcal{H}}^{EA}$, the following concentration bound holds*

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon} \geq \alpha\right\} \leq \exp\left(\frac{-c_{\boldsymbol{\mathcal{H}}^{EA}}\alpha^2 M\epsilon}{32S_{EA}^2K^4}\right)$$

Proof. Consider the random variable Θ (with randomness coming from the random initial condition distributed $\boldsymbol{\mu}^{\boldsymbol{Y}}$), and to ease the notation let $\widehat{\boldsymbol{\phi}}^{EA} := \widehat{\boldsymbol{\phi}}^{EA}_{L,\infty,\boldsymbol{\mathcal{H}}^{EA}}$,

$$\Theta = \frac{1}{L}\sum_{l=1}^{L}\left(\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}^{EA})\right)$$

The coercivity condition given in Definition (2.4.1), Proposition 2.8.5 and (2.8.4) allow us to bound the variance, denoted σ^2, of Θ as follows.

$$\begin{aligned}
\sigma^2 &\leq \mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\left|\frac{1}{L}\sum_{l=1}^{L}\left(\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}^{EA})\right)\right|^2\right] \\
&\leq \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left[\left|\boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}^{EA}_{\boldsymbol{X}(t_l)}(\widehat{\boldsymbol{\phi}}^{EA})\right|^2\right] \\
&\leq K^4\max\{R, R_{\dot{x}}\}^2\|\boldsymbol{\varphi}^{EA} - \widehat{\boldsymbol{\phi}}^{EA}\|^2_{\boldsymbol{L}^2(\boldsymbol{\rho}_T^L)}\|\boldsymbol{\varphi}^{EA} + \widehat{\boldsymbol{\phi}}^{EA} - 2\boldsymbol{\phi}^{EA}\|^2_\infty
\end{aligned}$$

$$\leq \frac{K^4 \max\{R, R_{\dot{x}}\}^2}{c_{\boldsymbol{\mathcal{H}}^{EA}}} (\boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}^{EA})) \|\boldsymbol{\varphi}^{EA} + \widehat{\boldsymbol{\phi}}^{EA} - 2\boldsymbol{\phi}^{EA}\|_\infty^2$$
$$\leq \frac{16 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4}{c_{\boldsymbol{\mathcal{H}}^{EA}}} (\boldsymbol{\mathcal{E}}_\infty^{EA}(\boldsymbol{\varphi}^{EA}) - \boldsymbol{\mathcal{E}}_\infty^{EA}(\widehat{\boldsymbol{\phi}}^{EA}))$$
$$\leq \frac{16 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4}{c_{\boldsymbol{\mathcal{H}}^{EA}}} \mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}). \tag{2.8.23}$$

By applying equation (2.8.9) from the proof of Proposition 2.8.1, we have that $\Theta \leq 8 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4$ almost surely. We then apply the one-sided Bernstein inequality to Θ and recalling the definitions (2.8.1) together with the definitions of the normalized errors in (2.8.21, 2.8.22), we get that:

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left\{ \frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon} \geq \alpha \right\} \leq \exp\left(- \frac{\alpha^2 (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)^2 M}{2\left(\sigma^2 + \frac{8 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4 \alpha (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)}{3} \right)} \right).$$

Now we provide a lower bound for the exponent to simplify the dependencies. We show that,

$$\frac{\epsilon \cdot c_{\boldsymbol{\mathcal{H}}^{EA}}}{32 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^6} \leq \frac{(\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)^2}{2\left(\sigma^2 + \frac{8 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4 \alpha (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)}{3} \right)},$$

or equivalently,

$$\frac{\epsilon \cdot c_{\boldsymbol{\mathcal{H}}^{EA}}}{16 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^6} \left(\sigma^2 + \frac{8 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^4 \alpha (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)}{3} \right)$$
$$\leq (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)^2 .$$

By the estimate (2.8.23), since $0 < \alpha \leq 1$, and $0 < c_{L,N,\boldsymbol{\mathcal{H}}^{EA}} < K^2$ it is sufficient to show that

$$\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})\epsilon + \frac{\epsilon(\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)}{6} \leq (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)^2.$$

This follows from Young's inequality as $2\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})\epsilon + \epsilon^2 \leq (\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon)^2$, and together these results give the desired bound of the proposition. □

We can easily derive the desired supremum bound by a covering argument. The estimation of the covering numbers involved will play a critical role in the main theorems and will be done in a dimension dependent way in order to get optimal minimax rates.

Proposition 2.8.11. *In the notation of Proposition 2.8.10,*

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\sup_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}}\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})-\mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})+\epsilon}\geq 3\alpha\right\}$$
$$\leq \mathcal{N}\left(\boldsymbol{\mathcal{H}}^{EA},\frac{\alpha\epsilon}{8S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}\right)e^{-\frac{c_{\boldsymbol{\mathcal{H}}^{EA}}\alpha^2M\epsilon}{32S_{EA}K^4}}$$

where $\mathcal{N}\left(\boldsymbol{\mathcal{H}}^{EA},\frac{\alpha\epsilon}{8S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}\right)$ *denotes the covering number of* $\boldsymbol{\mathcal{H}}^{EA}$ *with radius* $\frac{\alpha\epsilon}{8S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}$.

Proof. Let $\boldsymbol{\varphi}_i^{EA}=\boldsymbol{\varphi}_i^E\oplus\boldsymbol{\varphi}_i^A\in\boldsymbol{\mathcal{H}}^{EA}$, for $i=1,\ldots,\mathcal{N}\left(\boldsymbol{\mathcal{H}}^{EA},\frac{\alpha\epsilon}{8S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}\right)$, denote the center of disks D_i of radius $\frac{\alpha\epsilon}{8S_{EA}\max\{R,R_{\dot{x}}\}^2K^4}$ covering $\boldsymbol{\mathcal{H}}^{EA}$. The covering number is finite by the compactness assumption on the hypothesis space. By Lemma 2.8.8,

$$\sup_{\boldsymbol{\varphi}^{EA}\in D_i}\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})-\mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})+\epsilon}\geq 3\alpha\Rightarrow\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_i^{EA})-\mathcal{D}_M(\boldsymbol{\varphi}_i^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_i^{EA})+\epsilon}\geq\alpha.$$

Now, by Proposition 2.8.10, for each i,

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\sup_{\boldsymbol{\varphi}^{EA}\in D_i}\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})-\mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})+\epsilon}\geq 3\alpha\right\}\leq P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}_i^{EA})-\mathcal{D}_M(\boldsymbol{\varphi}_i^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}_i^{EA})+\epsilon}\geq\alpha\right\}$$
$$\leq e^{-\frac{c_{\boldsymbol{\mathcal{H}}^{EA}}\alpha^2M\epsilon}{32S_{EA}^2\max\{R,R_{\dot{x}}\}^2K^6}}.$$

By definition, $\boldsymbol{\mathcal{H}}^{EA}\subseteq\bigcup_i D_i$, so that

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}}\left\{\sup_{\boldsymbol{\varphi}^{EA}\in\boldsymbol{\mathcal{H}}^{EA}}\frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})-\mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA})+\epsilon}\geq 3\alpha\right\}$$

$$\leq \sum_i P_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left\{ \sup_{\boldsymbol{\varphi}^{EA} \in D_i} \frac{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) - \mathcal{D}_M(\boldsymbol{\varphi}^{EA})}{\mathcal{D}_\infty(\boldsymbol{\varphi}^{EA}) + \epsilon} \geq 3\alpha \right\}$$

$$\leq \mathcal{N}\left(\boldsymbol{\mathcal{H}}^{EA}, \frac{\alpha\epsilon}{8 S_{EA} \max\{R, R_{\dot{x}}\}^2 K^4} \right) e^{-\frac{c_{\boldsymbol{\mathcal{H}}^{EA}} \alpha^2 M\epsilon}{32 S_{EA}^2 \max\{R, R_{\dot{x}}\}^2 K^6}}.$$

□

Finally, we state the results for the ξ variable, the proofs are analogous. The advantage of splitting the theorems will become apparent. Specifically, it allows us to control the covering numbers on the EA and ξ hypothesis spacees separately, enabling us to get faster rates than if we viewed the task as estimating all functions simultaneously. This is possible due to the fundamentally decoupled nature of the dynamical system.

Proposition 2.8.12. *For all $\epsilon > 0$, $0 < \alpha < 1$, $\boldsymbol{\varphi}^{\xi} \in \boldsymbol{\mathcal{H}}^{\xi}$, the following concentration bound holds*

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left\{ \frac{\mathcal{D}_\infty^{\xi}(\boldsymbol{\varphi}^{\xi}) - \mathcal{D}_M^{\xi}(\boldsymbol{\varphi}^{\xi})}{\mathcal{D}_\infty^{\xi}(\boldsymbol{\varphi}^{\xi}) + \epsilon} \geq \alpha \right\} \leq e^{-\frac{c_{\boldsymbol{\mathcal{H}}^{\xi}} \alpha^2 M\epsilon}{32 S_0^2 K^4}}.$$

Proposition 2.8.13. *In the notation of Proposition 2.8.12,*

$$P_{\boldsymbol{\mu}^{\boldsymbol{Y}}} \left\{ \sup_{\boldsymbol{\varphi}^{\xi} \in \boldsymbol{\mathcal{H}}^{\xi}} \frac{\mathcal{D}_\infty^{\xi}(\boldsymbol{\varphi}^{\xi}) - \mathcal{D}_M^{\xi}(\boldsymbol{\varphi}^{\xi})}{\mathcal{D}_\infty^{\xi}(\boldsymbol{\varphi}^{\xi}) + \epsilon} \geq 3\alpha \right\} \leq \mathcal{N}\left(\boldsymbol{\mathcal{H}}^{\xi}, \frac{\alpha\epsilon}{8 S_0 R_\xi^2 K^4} \right) e^{-\frac{c_{\boldsymbol{\mathcal{H}}^{\xi}} \alpha^2 M\epsilon}{32 S_0 K^4}},$$

where $\mathcal{N}\left(\boldsymbol{\mathcal{H}}^{\xi}, \frac{\alpha\epsilon}{8 S_0 R_\xi^2 K^4} \right)$ denotes the covering number of $\boldsymbol{\mathcal{H}}^{\xi}$ with radius $\frac{\alpha\epsilon}{8 S_0 R_\xi^2 K^4}$.

2.9 Verification of coercivity condition

In this appendix, we study the coercivity condition for the second-order system of the form:

$$\begin{cases} m_i \ddot{\boldsymbol{x}}_i &= \mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N} (\phi^E(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\boldsymbol{x}_{i'} - \boldsymbol{x}_i) + \phi^A(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\dot{\boldsymbol{x}}_{i'} - \dot{\boldsymbol{x}}_i) \\ \dot{\xi}_i &= \mathbf{F}^{\xi}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N} \phi^{\xi}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\xi_{i'} - \xi_i) \end{cases} \tag{2.9.1}$$

We prove the coercivity condition for the system (2.9.1) in the case of $L = 1$.

When the system does not have a ξ variable, the system (2.9.1) is related to the anticipation dynamics studied in [121]. When $\phi^A \equiv 0$ or $\phi^E \equiv 0$, the system is called energy-based or alignment-based respectively, and has found application in various disciplines including opinion dynamics, particle dynamics, fish-milling dynamics, Cucker-Smale flocking dynamics and phototaxix dynamics. We refer the reader to [89, 90, 146] where extensive numerical experiments were conducted to demonstrate the effectiveness of the proposed learning approach on the aforementioned dynamics.

For conciseness, we only present the proof of coercivity for learning of ϕ^E and ϕ^A. A similar argument can be conducted to prove the coercivity for learning ϕ^{ξ}. Our aguments also work for special cases when $\phi^A \equiv 0$ or $\phi^E \equiv 0$. Therefore we obtain strict generalization of the coercivity results in [89, 90] which are only for first-order energy-based systems. We may go further to analyze the coercivity condition for heterogeneous systems. Compared to the homogeneous system, the coercivity condition would impose constraints on the angle between components of subspaces defined on the directed sum of measures, suggesting that the learning task is more difficult in the case $K \geq 2$.

Theorem 2.9.1. *Consider the system* (2.9.1) *at time* $t_1 = 0$ *with the initial distribution*

$\mu_0^{\boldsymbol{Y}} = \begin{bmatrix} \mu_0^{\boldsymbol{X}} \\ \mu_0^{\boldsymbol{V}} \\ \mu_0^{\boldsymbol{\Xi}} \end{bmatrix}$, *where* $\mu_0^{\boldsymbol{X}}$ *is exchangeable Gaussian with* $\mathrm{cov}(\boldsymbol{x}_i(t_1))-\mathrm{cov}(\boldsymbol{x}_i(t_1),\boldsymbol{x}_j(t_1)) = \lambda I_d$ *for a constant* $\lambda > 0$, $\mu_0^{\dot{\boldsymbol{X}}}, \mu_0^{\boldsymbol{\Xi}}$ *are exchangeable with finite second moment, and they are independent of* $\mu_0^{\boldsymbol{X}}$. *Then*

$$\mathbb{E}_{\mu_0^{\boldsymbol{Y}}} \|\mathbf{f}_{\varphi^E \oplus \varphi^A}(\boldsymbol{X}_0, \boldsymbol{V}_0)\|_{\mathcal{S}}^2 \geq c_{1,N,\mathcal{H}^{EA}} \left\|\varphi^E \oplus \varphi^A\right\|_{L^2(\rho_T^{EA,1})},$$
$$\mathbb{E}_{\mu_0^{\boldsymbol{Y}}} \|\mathbf{f}_{\varphi^\xi}(\boldsymbol{X}_0, \boldsymbol{\Xi}_0)\|_{\mathcal{S}}^2 \geq c_{1,N,\mathcal{H}^{\xi}} \left\|\varphi^\xi\right\|_{L^2(\rho_T^{\xi,1})},$$

(1) where we have

$$\rho_T^{EA,1}(r,\dot{r}) = \mathbb{E}_{\mu_{\mathbf{0}}^{\boldsymbol{Y}}}\left[\delta_{r_{12}(t_1),\dot{r}_{12}(t_1)}(r,\dot{r})\right] = \mathbb{E}_{\mu_0^{\boldsymbol{X}}}\delta_{r_{12}}(t_1)\mathbb{E}_{\mu_0^{\dot{\boldsymbol{X}}}}\delta_{\dot{r}_{12}}(t_1),$$
$$\rho_T^{\xi,1}(r,\xi) = \mathbb{E}_{\mu_{\mathbf{0}}^{\boldsymbol{Y}}}\left[\delta_{r_{12},\xi_{12}}(t_1)\right]$$
$$\left\|\varphi^E \oplus \varphi^A\right\|^2_{L^2(\rho_T^{EA,1})} = \|\varphi^E(r)r + \varphi^A(r)\dot{r}\|^2_{L^2(\rho_T^{EA,1}(r,\dot{r}))},$$
$$\left\|\varphi^\xi\right\|^2_{L^2(\rho_T^{\xi,1})} = \|\varphi^\xi(r)\xi\|^2_{L^2(\rho_T^{EA,1}(r,\xi))},$$

(2) $c_{1,N,\mathcal{H}^{EA}} \geq \frac{N-1}{2N^2} + \frac{(N-1)(N-2)}{2N^2}c$, $c = \min\left\{c^E_{\mathcal{H}^{EA}}, c^A_{\mathcal{H}^{EA}} c_{\mu_0^{\dot{\boldsymbol{X}}}}\right\}$ *with* $c_{\boldsymbol{\mu}_0^{\dot{\boldsymbol{X}}}} = 1 - \frac{\mathbb{E}\langle \dot{\boldsymbol{x}}_i(0), \dot{\boldsymbol{x}}_{i'}(0)\rangle}{\mathbb{E}\|\dot{\boldsymbol{x}}_i(0)\|^2}$ *($i \neq i'$) and* $c^E_{\mathcal{H}^{EA}}$ *and* $c^A_{\mathcal{H}^{EA}}$ *are non-negative constants and are positive for compact* $\mathcal{H}^{EA}$ *of* $L^2(\rho_T^{EA,1})$ *and independent of* N.

(3) $c_{1,N,\mathcal{H}^{\xi}} \geq (\frac{N-1}{N^2} + \frac{(N-1)(N-2)}{N^2}c), c = c_{\mathcal{H}^\xi} c_{\mu_0^{\boldsymbol{\Xi}}}$ *with* $c_{\mu_0^{\boldsymbol{\Xi}}} = 1 - \frac{\mathbb{E}\langle \xi_i(0), \xi_{i'}(0)\rangle}{\mathbb{E}\|\xi_i(0)\|^2}$ *($i \neq i'$) and* $c_{\mathcal{H}^\xi}$ *is a non-negative constant and is positive for compact* $\mathcal{H}^{\xi}$ *of* $L^2(\rho_T^{\xi,1})$ *and independent of* N.

Proof. The proof of part (1) follows from the definition of measures, norms and the

properties of the initial distributions. For part (2), we have

$$\begin{aligned}\mathbb{E}_{\mu_0^{\boldsymbol{Y}}}\|\mathbf{f}_{\varphi^E\oplus\varphi^A}(\boldsymbol{X}_0,\boldsymbol{V}_0)\|_{\mathcal{S}}^2 &= \frac{1}{N^3}\sum_{i=1}^{N}\left(\left(\sum_{j=k=1}^{N}+\sum_{j\neq k=1}^{N}\right)C_{i,j,k}^E+C_{i,j,k}^A+D_{i,j,k}\right)\\ &= \frac{N-1}{N^2}(\|\varphi^E(r)r\|^2_{L^2(\rho_T^{EA,1})}+\|\varphi^A(r)\dot r\|^2_{L^2(\rho_T^{EA,1})})+\mathcal{R}\\ &\geq \frac{N-1}{2N^2}\|\varphi^E\oplus\varphi^A\|^2_{L^2(\rho_T^{EA,1})}+\mathcal{R} \qquad (2.9.2)\end{aligned}$$

where

$$\begin{aligned}C_{i,j,k}^E &= \mathbb{E}_{\mu_0^{\boldsymbol{Y}}}\varphi^E(\|\boldsymbol{r}_{ji}(0)\|)\varphi^E(\|\boldsymbol{r}_{ki}(0)\|)\langle\boldsymbol{r}_{ji}(0),\boldsymbol{r}_{ki}(0)\rangle,\\ C_{i,j,k}^A &= \mathbb{E}_{\mu_0^{\boldsymbol{Y}}}\varphi^A(\|\boldsymbol{r}_{ji}(0)\|)\varphi^A(\|\boldsymbol{r}_{ki}(0)\|)\langle\dot{\boldsymbol{r}}_{ji}(0),\dot{\boldsymbol{r}}_{ki}(0)\rangle,\\ D_{i,j,k} &= \mathbb{E}_{\mu_0^{\boldsymbol{Y}}}\big(\varphi^E(\|\boldsymbol{r}_{ji}(0)\|)\varphi^A(\|\boldsymbol{r}_{ki}(0)\|)\langle\boldsymbol{r}_{ji}(0),\dot{\boldsymbol{r}}_{ki}(0)\rangle\\ &\qquad+\varphi^A(\|\boldsymbol{r}_{ji}(0)\|)\varphi^E(\|\boldsymbol{r}_{ki}(0)\|)\langle\dot{\boldsymbol{r}}_{ji}(0),\boldsymbol{r}_{ki}(0)\rangle\big)=0,\\ \mathcal{R} &= \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j\neq k,j\neq i,k\neq i}(C_{ijk}^A+C_{ijk}^E).\end{aligned}$$

By the property of $\mu_0^{\boldsymbol{Y}}$, we have

$$\begin{aligned}C_{ijk}^E &= \mathbb{E}\big[\varphi^E(\|X_1-X_2\|)\varphi^E(\|X_1-X_3\|)\,\langle X_2-X_1,X_3-X_1\rangle\,\big]\\ C_{ijk}^A &= \mathbb{E}\big[\varphi^A(\|X_1-X_2\|)\varphi^A(\|X_1-X_3\|)\big]\mathbb{E}\big[\,\langle Y_2-Y_1,Y_3-Y_1\rangle\,\big],\end{aligned}$$

for all (i,j,k), where X_is are exchangeable Gaussian random vectors with $\mathrm{cov}(X_1)-\mathrm{cov}(X_1,X_2)=\lambda I_d$ and Y_is are exchangeable random vectors who have the same distribution with the initial velocities of agents $\dot{\boldsymbol{x}}_i$s. From the Lemma 10 in [90] and Lemma 2.9.2 below,

$$\begin{aligned}C_{ijk}^E &\geq c_{\mathcal{H}^{EA}}^E\|\varphi^E\oplus 0\|^2_{L^2(\rho_T^{EA,1})}\\ C_{ijk}^A &\geq c_{\mathcal{H}^{EA}}^A c_{\mu_0^{\dot{\boldsymbol{X}}}}\|0\oplus\varphi^A\|^2_{L^2(\rho_T^{EA,1})},\qquad c_{\mu_0^{\dot{\boldsymbol{X}}}}=(1-\frac{\mathbb{E}\langle Y_1,Y_2\rangle}{\mathbb{E}\|Y_1\|^2}),\end{aligned}$$

where the constants $c^E_{\mathcal{H}^{EA}}, c^A_{\mathcal{H}^{EA}} \geq 0$ are always positive and independent of N for compact $\mathcal{H}^{EA}$, and we used the fact

$$\mathbb{E}\big[\, \langle Y_2 - Y_1, Y_3 - Y_1 \rangle \,\big] = \mathbb{E}\|Y_1\|^2 (1 - \frac{\mathbb{E}\langle Y_1, Y_2 \rangle}{\mathbb{E}\|Y_1\|^2}) \geq 0.$$

Therefore,

$$\mathbb{E}_{\mu_0}[\big\|\mathbf{f}_\varphi(\boldsymbol{X}_0, \dot{\boldsymbol{X}}_0)\big\|^2_{\mathcal{S}}] \geq c_{1,N,\mathcal{H}^{EA}} \|\varphi^E \oplus \varphi^A\|^2_{L^2(\rho^1_T(r,\dot{r}))},$$
$$c_{1,N,\mathcal{H}^{EA}} \geq \frac{N-1}{2N^2} + \frac{(N-1)(N-2)}{2N^2} \min\left\{ c^E_{\mathcal{H}^{EA}}, c^A_{\mathcal{H}^{EA}} c_{\mu_0^{\dot{\boldsymbol{X}}}} \right\}.$$

For part (3), the proof follows a similar path as for part (2). □

From Theorem 2.4.4, we see a particular case when the coercivity constant $c_{1,N,\mathcal{H}^{EA}}$ is positive uniformly in N if $c_{\mu_0^{\boldsymbol{V}}} > 0$. In fact, many distributions on $\mathbb{R}^{dN}$ with non-i.i.d $\mathbb{R}^d$ components make the constant $c_{\mu_0^{\boldsymbol{V}}}$ positive. For example, the components of $\boldsymbol{V}$ are exchangeable Gaussian but not i.i.d, and $d \geq 2$. In this particular case, coercivity is a property also of the system in the limit as $N \to \infty$, satisfying the mean-field equations. As a result, the estimation error of our estimators is independent of N.

The proof of Theorem 2.4.4 uses the following lemma, whose proof is the same as the proof of Lemma 10 in [90]. To be self-contained, we list the statement here.

Lemma 2.9.2. *Let X_1, X_2, X_3 be exchangeable Gaussian random vectors in $\mathbb{R}^d$ with $\mathrm{cov}(X_1) - \mathrm{cov}(X_1, X_2) = \lambda I_d$ for a constant $\lambda > 0$.*

- *The marginal distribution of $\rho_T^{EA,1}(r, \dot{r})$ with respect to r, denoted by $\rho(r)$, is a probability measure over $\mathbb{R}^+$ with density function $C_\lambda^{-1} r^{d-1} e^{-\frac{1}{4\lambda} r^2}$ where $C_\lambda = \frac{1}{2}(4\lambda)^{\frac{d}{2}} \Gamma(\frac{d}{2})$.*

- *We have*

$$\mathbb{E}\left[\varphi(|X_1 - X_2|)\varphi(|X_1 - X_3|)\right] \geq c_{\mathcal{X}} \|\varphi\|^2_{L^2(\rho)} \tag{2.9.3}$$

for all $\varphi \in \mathcal{X} \subset L^2(\rho)$*, with* $c_{\mathcal{X}} > 0$ *if* $\mathcal{X}$ *is compact and* $c_{\mathcal{X}} = 0$ *if* $\mathcal{X} = L^2(\rho)$*.*

2.10 Existence, uniqueness and properties of the measures

In this section, we provide technical details of the analytic properties of the collective system under consideration as well as of the measures that we defined in section 2.4.1. We emphasize that for the analytic portion of the theory, as we saw with the trajectory prediction result, we view the system (2.2.2) as coupled (whereas for the learning theory we leverage that they can be decoupled to make the estimation have better performance). We begin by showing that under the assumption that the interaction kernels lie in the corresponding admissible spaces, then the system is well-posed.

2.10.1 Well-posedness of second-order heterogeneous systems

Proposition 2.10.1. *Suppose the interaction kernels* $\boldsymbol{\phi}^E = (\phi^E_{kk'})^{K,K}_{k,k'=1}$*,* $\boldsymbol{\phi}^A = (\phi^A_{kk'})^{K,K}_{k,k'=1} \boldsymbol{\phi}^\xi = (\phi^\xi_{kk'})^{K,K}_{k,k'=1}$ *lie in the admissible sets* $\boldsymbol{\mathcal{K}}^E_{S_E}$,$\boldsymbol{\mathcal{K}}^A_{S_A}$, $\boldsymbol{\mathcal{K}}^\xi_{S_\xi}$ *respectively. Where the admissible spaces are defined in (2.3.11). Then the second-order heterogenous system* (2.2.2) *admits a unique global solution in* $[0, T]$ *for every initial datum* $\boldsymbol{X}_0, \dot{\boldsymbol{X}}(0) \in \mathbb{R}^{dN}$*,* $\boldsymbol{\Xi}(0) \in \mathbb{R}^N$ *and the solution depends continuously on the initial condition.*

The proof of Proposition 2.10.1 uses Lemma 2.10.2 and similar techniques used to prove the well-posedness of the first-order homogeneous system (see Section 6 in [19]) by rewriting the second-order system as a first-order system and then applying standard Caratheodory ODE results.

Lemma 2.10.2. *For any $\varphi^E \in \mathcal{K}^E_{S_E}$, $\varphi^A \in \mathcal{K}^A_{S_A}$, the function*

$$F[\varphi^{EA}](\boldsymbol{x}, \dot{\boldsymbol{x}}, s^E, s^A) := \varphi^E(||\boldsymbol{x}||, s^E)\boldsymbol{x} + \varphi^A(||\boldsymbol{x}||, s^A)\dot{\boldsymbol{x}},$$

for $\boldsymbol{x}, \dot{\boldsymbol{x}} \in \mathbb{R}^d$ is Lipschitz continuous on $\mathbb{R}^{2d+p^E+p^A}$ where p^E, p^A are the dimensions of the range of the functions s^E, s^A, respectively. Additionally, for any $\varphi^\xi \in \mathcal{K}^\xi_{S_\xi}$, the function

$$F[\varphi^\xi](\boldsymbol{x}, s^\xi, \xi) := \varphi^\xi(||\boldsymbol{x}||, s^\xi)\xi$$

is Lipschitz continuous on $\mathbb{R}^{d+1+p^\xi}$, where p^ξ is the dimension of the range of s^ξ.

2.10.2 Properties of measures

In this section we state and prove some technical properties of the measures described in Section 2.4.1.

Lemma 2.10.3. *Suppose each of the interaction kernels lie in the respective admissible spaces, namely, $\boldsymbol{\phi}^E \in \boldsymbol{\mathcal{K}}^E_{S_E}$,$\boldsymbol{\phi}^A \in \boldsymbol{\mathcal{K}}^A_{S_A}$, $\boldsymbol{\phi}^\xi \in \boldsymbol{\mathcal{K}}^\xi_{S_\xi}$. Then, for each (k, k'), the measures $\rho_T^{EA,k,k'}, \rho_T^{EA,L,k,k'}$ and $\rho_T^{\xi,k,k'}, \rho_T^{\xi,L,k,k'}$ defined in section 2.4.1, are regular Borel probability measures. Furthermore, if $\boldsymbol{\mu}^{\boldsymbol{Y}}$ is absolutely continuous with respect to the Lebesgue measure, then for each (k, k') we have that $\rho_T^{EA,k,k'}, \rho_T^{EA,L,k,k'}, \rho_T^{\xi,k,k'}, \rho_T^{\xi,L,k,k'}$ are absolutely continuous with respect to the Lebesgue measure. This implies Borel regularity, and under the absolute continuity of $\boldsymbol{\mu}^{\boldsymbol{Y}}$, absolute continuity with respect to Lebesgue measure of the measures, $\boldsymbol{\rho}_T^{EA}, \boldsymbol{\rho}_T^{\xi}, \boldsymbol{\rho}_T^{EA,L}, \boldsymbol{\rho}_T^{\xi,L}$.*

Proposition 2.10.4. *Suppose the distribution $\boldsymbol{\mu}^{\boldsymbol{Y}}$ of the initial condition is compactly supported. Then for each (k, k'), the support of the measures $\rho_T^{EA,kk'}, \rho_T^{\xi,kk'}$ (and therefore $\rho_T^{EA,L,kk'}, \rho_T^{\xi,L,kk'}$) is also compact.*

Proof. The compact support of the variables $r, \dot{r}, \xi$ and the feature maps follows by the global well-posedness of the system in finite time, together with the Lipschitz

assumptions on the non-collective forces. This compact support over a fixed, finite time is what is claimed in Proposition 2.10.4. □

The main point is that by making reasonable assumptions on the non-collective forces, feature maps, interaction kernels, and the interval of time, together with the assumption that our agents' initial conditions cannot be arbitrarily far apart, we can derive that the pairwise distance, velocity and ξ will be controlled. Thus, the measures in section 2.4.1, if given enough trajectories, will be well approximated by the discretized version using the numerical approach described in section 2.5. Meaning that if we have a reasonable number of trajectories, we can look at the set of pairwise distances, velocities, etc. that these agents explore and bin them to set the support of the interaction kernels. Explicit values for the constants claimed in the proposition depend on the properties of the non-collective forces, the support and sup-norm of the interaction kernels, the interval T, and the number of agents.

2.11 Background results

In this section, for the convenience of the reader, we gather a few of the technical tools used in the analysis of the system. These are fundamental results necessary for developing the trajectory prediction, the measure support, and the existence and uniqueness results. We also include some of the necessary results on covering numbers of function spaces used for the learning theory.

The first theorem we present is an iterated Grönwall type result that allows us to analyze the trajectory error of the full system $\boldsymbol{Y}(t)$.

Theorem 2.11.1. *Let $u(t), a(t),$ and $b(t)$ be nonnegative continuous functions in*

$J = [\alpha, \beta]$, and suppose that

$$\begin{aligned} u(t) \leq & a(t) + b(t) \left[\int_\alpha^t k_1(t, t_1) u(t_1) \, dt_1 + \cdots \right. \\ & \left. + \int_\alpha^t \left(\int_\alpha^{t_1} \cdots \left(\int_\alpha^{t_{n-1}} k_n(t, t_1, \ldots, t_n) u(t_n) \, dt_n \right) \cdots \right) dt_1 \right] \end{aligned}$$

for all $t \in J$, where $k_i(t, t_1, \ldots, t_i)$ are nonnegative continuous functions in $J_{i+1}, i = 1, 2, \ldots, n$, which are nondecreasing in $t \in J$ for all fixed $(t_1, \ldots, t_i) \in J_i, i = 1, 2, \ldots, n$. Then, for all $t \in J$

$$u(t) \leq a(t) + b(t) \int_\alpha^t \widehat{R}[a](t, s) \exp\left(\int_s^t \widehat{R}[b](t, \tau) d\tau \right) ds$$

where, for all $(t, s) \in J_2$

$$\begin{aligned} \widehat{R}[w](t, s)) = & k_1(t, s) w(s) + \int_\alpha^s k_2(t, s, t_2) w(t_2) \, dt_2 \\ & + \sum_{i=3}^n \int_\alpha^s \left(\int_\alpha^{t_2} \cdots \left(\int_\alpha^{t_{t-1}} k_i(t, s, t_2, \ldots, t_i) w(t_i) \, dt_i \right) \cdots \right) dt_2 \end{aligned}$$

for each continuous function $w(t)$ in J.

Proof. See [34]. □

2.12 Additional comments on first-order models and theory

Our second-order model formulation covers the first-order equations of [90, 89] as a special case. When $\mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) = -\nu_i \dot{\boldsymbol{x}}_i + \mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \xi_i)$ for some constant $\nu_i > 0$, $\mathbf{F}^{\xi}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) = \mathbf{F}^{\xi}(\boldsymbol{x}_i, \xi_i)$, $\phi^A_{k_i k_{i'}} \equiv 0$ for all $k, k' = 1, \ldots, K$, and $m_i \ll 1$, (2.2.2)

becomes,

$$\begin{cases} \nu_i \dot{\boldsymbol{x}}_i & = \mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^E_{k_i k_{i'}}(r_{ii'}, \boldsymbol{s}^E_{ii'})(\boldsymbol{x}_{i'} - \boldsymbol{x}_i) \\ \dot{\xi}_i & = \mathbf{F}^{\xi}(\boldsymbol{x}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^{\xi}_{k_i, k_{i'}}(r_{ii'}, \boldsymbol{s}^{\xi}_{ii'})(\xi_{i'} - \xi_i) \end{cases}$$

It extends the first-order models considered in [89, 90, 146] by adding non-collective forces, $\mathbf{F}^{\boldsymbol{x}}, \mathbf{F}^{\xi}$, multi-dimensional interaction kernels, $\phi^E_{k,k'}, \phi^{\xi}_{k,k'}$, and auxiliary variables, ξ_i.

The first-order theory considered in [89, 90] was focused on the learnability of functions of the form, $\phi^E(r)r$, which is a special case of our second-order theory, where we study the functions of the form $\phi^E(r)r + \phi^A(r)\dot{r}$, with $\phi^A(r) \equiv 0$.

2.13 Additional performance measures

For measures related to learning the ξ-based interaction kernels, we take

$$\begin{cases} \delta^{\xi}_{i,i',t}(r, \boldsymbol{s}^{\xi}, \xi) & := \delta_{r_{ii'}(t), \boldsymbol{s}^{\xi}_{ii'}(t), \xi_{ii'}(t)}(r, \boldsymbol{s}^{\xi}, \xi) \\ \delta^{\xi}_{i,i',t,m}(r, \boldsymbol{s}^{\xi}, \xi) & := \delta_{r^{(m)}_{ii'}(t), \boldsymbol{s}^{\xi,(m)}_{ii'}(t), \xi^{(m)}_{ii'}(t)}(r, \boldsymbol{s}^{\xi}, \xi) \end{cases}$$

Then, the measures are given by

$$\rho^{\xi,k,k'}_T(r, \boldsymbol{s}^{\xi}, \xi) := \mathbb{E}_{\boldsymbol{Y}_0 \sim \boldsymbol{\mu}^{\boldsymbol{Y}}} \frac{1}{T N_{kk'}} \int_{t=0}^{T} \sum_{\substack{i \in C_k, i' \in C_{k'} \\ i \neq i'}} \delta^{\xi}_{ii',t}(r, \boldsymbol{s}^{\xi}, \xi)\, dt \tag{2.13.1}$$

and similarly for $\rho^{\xi,L,k,k'}_T(r, \boldsymbol{s}^{\xi}, \xi)$, $\rho^{\xi,L,M,k,k'}_T(r, \boldsymbol{s}^{\xi}, \xi)$. Similarly, $\rho^{\xi,k,k'}_T$ and its time-discretization version, $\rho^{\xi,L,k,k'}_T$, are only used in the theoretical setting, whereas the empirical $\rho^{\xi,L,M,k,k'}_T$ is used in the actual algorithm. We consider direct sums of the

measures for the phase variable for ease of notation.

$$\boldsymbol{\rho}_T^{\xi,L} = \bigoplus_{k,k'=1,1}^{K,K} \rho_T^{\xi,L,kk'}, \quad \boldsymbol{\rho}_T^{\xi} = \bigoplus_{k,k'=1,1}^{K,K} \rho_T^{\xi,kk'}, \quad \boldsymbol{L}^2\left(\boldsymbol{\rho}_T^{\xi,L}\right) = \bigoplus_{k,k'=1,1}^{K,K} L^2\left(\rho_T^{\xi,L,kk'}\right) \tag{2.13.2}$$

Lastly, for the ξ-based interaction kernels, i.e., $\hat{\phi}_{kk'}^{\xi}$ versus $\phi_{kk'}^{\xi}$, we consider the following norm,

$$\left\|\hat{\phi}_{kk'}^{\xi} - \phi_{kk'}^{\xi}\right\|^2_{L^2(\rho_T^{\xi,k,k'})} = \int_r \int_{\xi} \int_{\boldsymbol{s}^{\xi}} (\hat{\phi}_{kk'}^{\xi}(r,\xi,\boldsymbol{s}^{\xi}) - \phi_{kk'}^{\xi}(r,\xi,\boldsymbol{s}^{\xi}))^2 \xi^2 \, d\rho_T^{\xi,k,k'}(r,\xi,\boldsymbol{s}^{\xi}). \tag{2.13.3}$$

2.14 Numerical algorithm

In this section, we will detail the construction of the linear systems to learn $\vec{\alpha}^{EA}$ and $\vec{\alpha}^{\xi}$.

We start with the procedure of solving for $\vec{\alpha}^{EA}$. First, we build the basis functions for the finite dimensional hypothesis spaces $\mathcal{H}_{kk'}^{E}, \mathcal{H}_{kk'}^{A}$ using piecewise polynomials or clamped B-splines as the basis functions (see 2.3.6), which altogether are represented as $\boldsymbol{\mathcal{H}}^{EA}$ as in (2.3.15).

Remark 2.14.1. *The support of the unknown interaction kernels is not assumed to be known. We build our finite dimensional subspaces, $\mathcal{H}_{kk'}^{E}, \mathcal{H}_{kk'}^{A}$, based on the empirical observation data. For the support-detection capability of our estimators, see the examples of opinion dynamics in [89, 146].*

We utilize the tensor grid of basis functions, i.e., tensor product of basis functions in each dimension of the basis $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}] \times \mathbb{S}_{kk'}^{E,L,M}$ or $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}] \times \mathbb{S}_{kk'}^{A,L,M}$, where $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}]$ is the empirical range of r given by the observation data, similarly for the empirical $\mathbb{S}_{kk'}^{E,L,M}$ and $\mathbb{S}^{A,L,M}$ being the range of $\boldsymbol{s}_{kk'}^{E}$ and $\boldsymbol{s}_{kk'}^{A}$ given

by the observation, respectively. In each dimension[4] of $[R^{\min,L,M}_{kk'}, R^{\max,L,M}_{kk'}] \times \mathbb{S}^{E,L,M}_{kk'}$ or $[R^{\min,L,M}_{kk'}, R^{\max,L,M}_{kk'}] \times \mathbb{S}^{A,L,M}_{kk'}$, the basis functions are built as piecewise standard polynomials (or other functions, such as Clamped B-splines, Fourier basis, etc.) uniformly with the number of basis functions being $n^{E,j}_{kk'}$ or $n^{A,j}_{kk'}$. Hence $n^{E}_{kk'} = \prod_j^{1+p^E_{k,k'}} n^{E,j}_{kk'}$ and $n^{A}_{kk'} = \prod_j^{1+p^A_{k,k'}} n^{A,j}_{kk'}$. Then, we assemble $\vec{d}^{(m)}$ as follows,

$$\vec{d}^{EA,(m)} = \begin{bmatrix} \frac{1}{\sqrt{N_{\ell_1}}}\ddot{\boldsymbol{x}}_1(t_1) \\ \vdots \\ \frac{1}{\sqrt{N_{\ell_N}}}\ddot{\boldsymbol{x}}_N(t_1) \\ \vdots \\ \frac{1}{\sqrt{N_{\ell_1}}}\ddot{\boldsymbol{x}}_1(t_L) \\ \vdots \\ \frac{1}{\sqrt{N_{\ell_N}}}\ddot{\boldsymbol{x}}_N(t_L) \end{bmatrix}.$$

If $\ddot{\boldsymbol{x}}_i(t_l)$ is not given, a finite difference scheme on $\boldsymbol{x}_i(t_l)$ or $\dot{\boldsymbol{x}}_i(t_l)$ is used to approximate $\ddot{\boldsymbol{x}}_i(t_l)$. Next, we build, $\vec{f}^{(m)}$ as follows,

$$\vec{f}^{EA,(m)} = \begin{bmatrix} \frac{1}{\sqrt{N_{\ell_1}}}\mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_1(t_1), \dot{\boldsymbol{x}}_1(t_1), \xi_1(t_1)) \\ \vdots \\ \frac{1}{\sqrt{N_{\ell_N}}}\mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_N(t_1), \dot{\boldsymbol{x}}_N(t_1), \xi_N(t_1)) \\ \vdots \\ \frac{1}{\sqrt{N_{\ell_1}}}\mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_1(t_L), \dot{\boldsymbol{x}}_1(t_L), \xi_1(t_L)) \\ \vdots \\ \frac{1}{\sqrt{N_{\ell_N}}}\mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_N(t_L), \dot{\boldsymbol{x}}_N(t_L), \xi_N(t_L)) \end{bmatrix}.$$

Then for the learning matrix, $\Psi^{EA,(m)} \in \mathbb{R}^{LNd \times n}$ with $n = n^E + n^A$. It is a concatnation

[4]Mixture of basis functions in each dimension is possible, the algorithm does not required the basis functions in each dimension to be of the same kind. We make such assumption for simplicity sake.

of two sub-matrix, $\Psi^{E,(m)}$ and $\Psi^{A,(m)}$, i.e.,

$$\Psi^{EA,(m)} = \begin{bmatrix} \Psi^{E,(m)} & \Psi^{A,(m)} \end{bmatrix}.$$

For the energy-based learning matrix, $\Psi^{E,(m)}$, we use a lexicographical order on (k,k') for $k,k' = 1,\dots,K$. We define $n^E_{k,k',\text{prev}} = \sum_{(k_1,k_2)<(k,k')} n^E_{k_1,k_2}$; if $(k,k') = (1,1)$, we take $n^E_{1,1,\text{prev}} = 0$. Then for $\eta^E_{kk'} = 1,\dots,n^E_{k,k'}$, $\Psi^{E,(m)}$ is given as follows, for $i \in C_k$,

$$\Psi^{E,(m)}(li(1:d),\eta^E_{kk'}) = \sum_{i' \in C_{k'}} \frac{1}{\sqrt{N_{k_i}}} \psi^{\boldsymbol{x}}_{k,k',\eta^E_{kk'}}(\|\boldsymbol{x}_{i'}(t_l) - \boldsymbol{x}_i(t_l)\|, \boldsymbol{s}^E_{i,i'}(t_l))(\boldsymbol{x}_{i'}(t_l) - \boldsymbol{x}_i(t_l)),$$

and for $l = 1,\dots,L$. Similar process of construction is done for $\Psi^{A,(m)}$. Then we define,

$$A^{EA,(m)} = (\Psi^{EA,(m)})^T \Psi^{EA,(m)} \quad \text{and} \quad \vec{b}^{EA,(m)} = (\Psi^{EA,(m)})^T (\vec{d}^{EA,(m)} - \vec{f}^{EA,(m)}).$$

And lastly,

$$A^{EA}_M = \frac{1}{LM} \sum_{m=1}^{M} A^{EA,(m)} \quad \text{and} \quad \vec{b}^{EA} = \frac{1}{LM} \sum_{m=1}^{M} \vec{b}^{EA,(m)}.$$

Then, $\vec{\alpha}^{EA} = \begin{bmatrix} (\vec{\alpha}^E)^T & (\vec{\alpha}^A)^T \end{bmatrix}^T$, is obtained by solving

$$A^{EA}_M \vec{\alpha}^{EA} = \vec{b}^{EA}.$$

Then, we assemble

$$\hat{\phi}^E_{kk'} = \sum_{\eta^E_{kk'}=1}^{n^E_{k,k'}} \alpha^E_{k,k',\eta^E_{kk'}} \psi^{\boldsymbol{x}}_{k,k',\eta^E_{kk'}}.$$

Similar assembly from α^A is done for $\hat{\phi}^A_{kk'}$. In the case of using finite difference

approximation to approximate the second derivatives of $\boldsymbol{x}_i$, we end up with

$$A_M^{EA}\vec{\alpha}^{EA} = \vec{b}^{EA} + \vec{\zeta},$$

where $\vec{\zeta} = \mathcal{O}(\frac{T}{L})$ when a first-order finite difference scheme is used.

Next for $\vec{\alpha}^{\xi}$, we build the basis functions for each of the finite dimensional spaces $\mathcal{H}_{kk'}^{\xi}$, using piecewise polynomials or clamped B-splines (as in the EA case, similarly, other many other bases work well in this algorithm). This is an explicit example of $\boldsymbol{\mathcal{H}}^{\xi}$. We utilize the tensor grid of basis functions, i.e., tensor product of basis functions in each dimension of the basis $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}] \times \mathbb{S}_{kk'}^{\xi,L,M}$, where $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}]$ is the empirical range of r given by the observation data, similarly for the empirical $\mathbb{S}_{kk'}^{\xi,L,M}$ being the range of $\boldsymbol{s}_{kk'}^{\xi}$ given by the observation. And in each dimension of $[R_{kk'}^{\min,L,M}, R_{kk'}^{\max,L,M}] \times \mathbb{S}_{kk'}^{\xi,L,M}$, the basis functions are built as piecewise standard polynomials (or other functions, such as Clamped B-splines, Fourier basis, etc.) uniformly with the number of basis functions being $n_{kk'}^{\xi,j}$. Hence $n_{kk'}^{\xi} = \prod_j^{1+p_{k,k'}^{\xi}} n_{kk'}^{\xi,j}$. We let

$$\vec{d}^{\xi,(m)} := \begin{bmatrix} \frac{1}{\sqrt{N_{k_1}}}\dot{\xi}_1(t_1) \\ \vdots \\ \frac{1}{\sqrt{N_{k_N}}}\dot{\xi}_N(t_1) \\ \vdots \\ \frac{1}{\sqrt{N_{k_1}}}\dot{\xi}_1(t_L) \\ \vdots \\ \frac{1}{\sqrt{N_{k_N}}}\dot{\xi}_N(t_L) \end{bmatrix}, \quad \vec{f}^{\xi,(m)} := \begin{bmatrix} \frac{1}{\sqrt{N_{k_1}}}\mathbf{F}^{\xi}(\boldsymbol{x}_1(t_1), \dot{\boldsymbol{x}}_1(t_1), \xi_1(t_1)) \\ \vdots \\ \frac{1}{\sqrt{N_{k_N}}}\mathbf{F}^{\xi}(\boldsymbol{x}_N(t_1), \dot{\boldsymbol{x}}_N(t_1), \xi_N(t_1)) \\ \vdots \\ \frac{1}{\sqrt{N_{k_1}}}\mathbf{F}^{\xi}(\boldsymbol{x}_1(t_L), \dot{\boldsymbol{x}}_1(t_L), \xi_1(t_L)) \\ \vdots \\ \frac{1}{\sqrt{N_{k_N}}}\mathbf{F}^{\xi}(\boldsymbol{x}_N(t_L), \dot{\boldsymbol{x}}_N(t_L), \xi_N(t_L)) \end{bmatrix},$$

and for $i \in C_k$

$$\Psi^{\xi,(m)}(li(1:d), \eta_{kk'}^{\xi}) = \sum_{i' \in C_{k'}} \frac{1}{\sqrt{N_{k_i}}} \psi_{k,k',\eta_{kk'}^{\xi}}^{\xi}(\|\boldsymbol{x}_{i'}(t_l) - \boldsymbol{x}_i(t_l)\|, \boldsymbol{s}_{i,i'}^{\xi}(t_l))(\xi_{i'}(t_l) - \xi_i(t_l)).$$

Finally we define,

$$A^{\xi,(m)} = (\Psi^{\xi,(m)})^T \Psi^{\xi,(m)} \quad \text{and} \quad \vec{b}^{\xi,(m)} = (\Psi^{\xi,(m)})^T (\vec{d}^{\xi,(m)} - \vec{f}^{\xi,(m)}).$$

and

$$A^{\xi} = \frac{1}{LM} \sum_{m=1}^{M} A^{\xi,(m)} \quad \text{and} \quad \vec{b}^{\xi} = \frac{1}{LM} \sum_{m=1}^{M} \vec{b}^{\xi,(m)}.$$

Thus, $\vec{\alpha}^{\xi}$ is obtained by solving

$$A^{\xi} \vec{\alpha}^{\xi} = \vec{b}^{\xi}.$$

Then, we assemble

$$\hat{\phi}^{\xi}_{kk'} = \sum_{\eta^{\xi}_{kk'}=1}^{n^{\xi}_{k,k'}} \alpha^{\xi}_{k,k',\eta^{\xi}_{kk'}} \psi^{\xi}_{k,k',\eta^{\xi}_{kk'}}.$$

Chapter 3

Emergent Behaviors

3.1 Introduction

Our work is focused on discovering governing equations of collective dynamics (also known as self-organized dynamics), a special kind of interacting particle- and agent-based dynamical systems. This chapter considers models that are a particular case of the models proposed in Chapter 2. The material in it is drawn from [146] and uses slightly different notation from that chapter – which we retain for historical accuracy as the work of this chapter was done before the work for Chapter 2 and in fact informed it. Our main goal in this chapter is to explore how the large-time and emergent behavior of these systems is captured by the ones driven by the estimated interaction kernels. The models we consider are a distinguished subset of general autonomous systems of ODEs

$$\dot{\boldsymbol{X}}(t) = \mathbf{f}(\boldsymbol{X}(t)), \quad \boldsymbol{X}(T_0) = \boldsymbol{X}_0 \in \mathbb{R}^D \quad \text{and} \quad t \in [0, T]. \tag{3.1.1}$$

In this general case, given $\{\boldsymbol{X}(t)\}_{t\in[T_0,T]}$ $(0 \le T_0 < T)$, system identification consists in inferring $\mathbf{f}$ from $\boldsymbol{X}(t)$ and $\dot{\boldsymbol{X}}(t)$ observed at various t's. Classical regression techniques (e.g. [132, 64, 55, 15]) have recently been brought to bear on this problem (e.g.

[118, 23, 129, 14]). However, lack of independence among the observation data and the curse of dimensionality due to D typically being large, are all obstructions to finding the desired $\mathbf{f}$ effectively and efficiently (see [89] for an extended discussion) of these points. The works in [19, 89, 90] proposed a learning approach that exploits the special form of collective dynamics to overcome the difficulties mentioned above. As we saw in Chapter 2, the study began with first order systems of the form

$$\dot{\boldsymbol{x}}_i = \frac{1}{N}\sum_{i'=1}^{N} \phi(||\boldsymbol{x}_{i'} - \boldsymbol{x}_i||)(\boldsymbol{x}_{i'} - \boldsymbol{x}_i) \quad , \quad i = 1, \dots, N\,. \tag{3.1.2}$$

The 1-dimensional function ϕ is referred to as the *interaction kernel*. Here and in what follows we assume, with possibly abuse of notation, that the term $i' = i$ in the sum in the r.h.s. is $\mathbf{0}$, even in cases where ϕ may not be defined at 0 (e.g. $\phi(r) = 1/r^2$). The aforementioned works consider the problem of estimating ϕ given trajectory observations, in terms of positions and velocities of the agents at various times, along one or multiple trajectories (with different initial conditions (ICs), e.g. sampled at random from some probability distribution on the state space). In [19, 89] a nonparametric learning approach to construct an estimator $\widehat{\phi}$ for ϕ is considered, that exploits the governing structure of the dynamics in (3.1.2), which is a special (yet ubiquitous) case of the general equations (3.1.1). The work [19] considered a first-order model of homogeneous agents (derived from gradient flow), and studied the convergence to its mean field limit and the inference of the mean-field limit interaction kernel from observations of trajectories of the system with a finite and yet increasing number of agents. The work [89] extended the approach in [19] to the situation where the number of agents is fixed, but the number of observations increases, showing that the nonparametric estimators for the interaction kernel converge at the near optimal rate for regression in one dimension, in particular independent of the dimension of the state space. It generalized the estimators to first and second-order of heterogeneous

agents with 1-dimensional interaction kernels based on pairwise distances, providing substantial numerical evidence of the performance of these generalizations. The work [90] analyzes in detail the estimators for first order heterogeneous agent-based systems, generalizing the theoretical results of [89] to that case, while sharpening some of the constructions. In chapter 2 we demonstrated a comprehensive theory for these systems and showed an optimal convergence rate, consistency, and other desirable statistical properties.

In this chapter, we will show that the estimated interaction kernels can also provide insight into discovering the correct emergent behaviors at large time, as we will demonstrate in several examples in section 3.5 and section 3.6. It should be noted that the theory of Chapter 2 does not have guarantees about the correctness of the emergent properties (which are often qualitative phenomena which we measure quantitatively) of the trajectories directly, albeit our guarantees on the accuracy of the estimators and trajectories helps to explain the strong performance of the estimators at the task of correctly identifying emergent phenomena in agent-based dyamical systems. We also consider examples from the heterogeneous case, as we did in Chapter 2, where we have a family of interaction kernels $\{\phi_k\}_k$. We consider the case of gravity in section 3.7, and show that we can discover both the "common structure" of the interaction kernel, namely the $1/r^2$ dependency on pairwise distance, and the dependency on mass.

The structure of this chapter is as follows. In section 3.2 we discuss in detail the two models, of first and second-order systems respectively, which we are considering. In section 3.3 we outline the learning algorithm for each model. These algorithms are efficient and scalable. They enjoy highly favorable performance in terms of computational complexity as described in section 3.3.1. Section 3.4 defines the various performance measures, confusion matrices, and pattern indicator scores. It also discuses how we set up the numerical experiments. Section 3.5 to Section 3.7

provide a detailed study of five fundamental dynamical systems that vary across order, interaction kernel form, and agent characteristics, as well as learning of interaction kernels that involve parameters. Finally, we conclude the chapter and discuss various future research directions stimulated by these results in section 3.8.

3.2 Model Description

The main focus of this work is to numerically investigate the capability of the estimators at predicting emergent behaviors of various collective dynamics using our extended learning approach. Our extended learning covers dynamical systems much more elaborate than those considered in [89]. These systems have a complex balance between non-collective and collective forces, include interaction laws depending on more than one variable, and allow for parametric families of interaction laws. We motivate these extensions with families of dynamical systems exhibiting these more intricate governing structures, which were motivated by various applications and whose dynamical properties have been studied in the scientific literature.

Here we consider particle- and agent-based systems that model rather general complex systems, beyond those considered in [19, 89, 90]. The first-order models are governed by the following system of coupled ODEs

$$\begin{cases} \dot{\boldsymbol{x}}_i &= \mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^E_{k_i, k_{i'}}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|, s^{\boldsymbol{x}}_{i,i'})(\boldsymbol{x}_{i'} - \boldsymbol{x}_i) \\ \dot{\xi}_i &= \mathbf{F}^{\xi}(\boldsymbol{x}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^{\xi}_{k_i, k_{i'}}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|, s^{\xi}_{i,i'}) \end{cases}, i = 1, \cdots, N \tag{3.2.1}$$

These systems contain heterogeneous agents: the agents are partitioned into K different types, with C_k containing the indices of the agents of type k, for $k = 1, \ldots, K$. Table 3.1 shows the definitions of variables in (3.2.1).

Variable	Definition
$\boldsymbol{x}_i = \boldsymbol{x}_i(t) \in \mathbb{R}^d$	state vector (positions, opinions, etc.)
$\xi_i = \xi_i(t) \in \mathbb{R}$	auxiliary variable (phase, headings, etc.)
N	number of agents
k_i	type index of agent i
N_k	number of agents in type k
$\|\cdot\|$	any norm on $\mathbb{R}^d$ (usually an ℓ_2 norm)
$\mathbf{F}^{\boldsymbol{x}}, \mathbf{F}^{\xi}$	non-collective changes on $\dot{\boldsymbol{x}}_i$ and $\dot{\xi}_i$, respectively
$\phi^E_{k_i,k_{i'}}, \phi^{\xi}_{k_i,k_{i'}}$	interaction kernels: how the agents in type $k_{i'}$ influence agents in type k_i
$s^{\boldsymbol{x}}_{i,i'}$	$\mathcal{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \xi_i, \boldsymbol{x}_{i'}, \xi_{i'}) : \mathbb{R}^{2d+2} \to \mathbb{R}$
$s^{\xi}_{i,i'}$	$\mathcal{F}^{\xi}(\boldsymbol{x}_i, \xi_i, \boldsymbol{x}_{i'}, \xi_{i'}) : \mathbb{R}^{2d+2} \to \mathbb{R}$

Table 3.1: Notation for first-order models

Remark 3.2.1. *Compared to Eqn.* (8) *in [89], our new equation* (3.2.1) *has the following additions: the new ξ_i variable, non-collective forces $\mathbf{F}^{\boldsymbol{x}}, \mathbf{F}^{\xi}$, and* 2*-dimensional interaction laws (as opposed to only single-variable, pairwise-distance-based interactions).*

The second-order models we consider are governed by the following system of coupled ODEs,

$$\begin{cases} m_i \ddot{\boldsymbol{x}}_i &= \mathbf{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \Big[\phi^E_{k_i,k_{i'}}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|, s^{\boldsymbol{x}}_{i,i'})(\boldsymbol{x}_{i'} - \boldsymbol{x}_i) \\ & \qquad\qquad + \phi^A_{k_i,k_{i'}}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|, s^{\dot{\boldsymbol{x}}}_{i,i'})(\dot{\boldsymbol{x}}_{i'} - \dot{\boldsymbol{x}}_i) \Big] \\ \dot{\xi}_i &= \mathbf{F}^{\xi}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i) + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^{\xi}_{k_i,k_{i'}}(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|, s^{\xi}_{i,i'})(\xi_{i'} - \xi_i) \end{cases} \tag{3.2.2}$$

for $i = 1, \cdots, N$. Table 3.2 shows the definitions of variables in (3.2.2). Natural regularity and growth assumptions on the functions in the right-hand side are made so that the system has a unique solution for all times. For example assuming that the functions involved are at least Lipschitz and decay sufficiently rapidly at infinity would suffice.

Variable	Definition
m_i	mass of agent i
$\mathbf{F}^{\dot{\boldsymbol{x}}}, \mathbf{F}^{\xi}$	non-collective changes on $\ddot{\boldsymbol{x}}_i$ and $\dot{\xi}_i$ respectively
$\phi^E, \phi^A, \phi^{\xi}$	energy, alignment, and environment-based interaction kernels respectively
$s^{\boldsymbol{x}}_{i,i'}$	$\mathcal{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i, \boldsymbol{x}_{i'}, \dot{\boldsymbol{x}}_{i'}, \xi_{i'}) : \mathbb{R}^{4d+2} \to \mathbb{R}$
$s^{\dot{\boldsymbol{x}}}_{i,i'}$	$\mathcal{F}^{\dot{\boldsymbol{x}}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i, \boldsymbol{x}_{i'}, \dot{\boldsymbol{x}}_{i'}, \xi_{i'}) : \mathbb{R}^{4d+2} \to \mathbb{R}$
$s^{\xi}_{i,i'}$	$\mathcal{F}^{\xi}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i, \xi_i, \boldsymbol{x}_{i'}, \dot{\boldsymbol{x}}_{i'}, \xi_{i'}) : \mathbb{R}^{4d+2} \to \mathbb{R}$

Table 3.2: Notation for second-order models

Remark 3.2.2. *Compared to Eqn.* (11) *in [89], our new equation* (3.2.1) *has the following additions: slightly different non-collective forces,* $\mathbf{F}^{\dot{\boldsymbol{x}}}, \mathbf{F}^{\xi}$*, and* 2*-dimensional interaction laws.*

We are given observation data, namely $\{\boldsymbol{y}_i^m(t_l), \dot{\boldsymbol{y}}_i^m(t_l)\}_{i,m=1}^{N,M}$ ($\boldsymbol{y}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \xi_i \end{bmatrix}$ for first order systems or $\boldsymbol{y}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \dot{\boldsymbol{x}}_i \\ \xi_i \end{bmatrix}$ for second order systems) at time instances $T_0 = t_1 < \cdots < t_L = T$. In the case of missing derivative data, namely $\dot{\boldsymbol{y}}_i^m$, we will approximate $\dot{\boldsymbol{y}}_i^m$ using appropriate finite difference schemes. The observation data is generated from M initial conditions (ICs), $\{(\boldsymbol{y}_i^m(0))_i\}_m$, which are i.i.d samples from a (typically unknown) probability distribution $\mu^{\boldsymbol{y}}$ ($\mu^{\boldsymbol{y}} = \mu^{\boldsymbol{x}} \oplus \mu^{\xi}$ for first order and $\mu^{\boldsymbol{y}} = \mu^{\boldsymbol{x}} \oplus \mu^{\dot{\boldsymbol{x}}} \oplus \mu^{\xi}$ for second order). The unknowns in these systems are the interaction laws and the distribution of the initial conditions, while everything else is assumed known. We construct estimators for $\phi^E_{k_i,k_{i'}}, \phi^{\xi}_{k_i,k_{i'}}$ (resp. $\phi^E_{k_i,k_{i'}}, \phi^A_{k_i,k_{i'}}, \phi^{\xi}_{k_i,k_{i'}}$ for second-order systems) that are close to the true interaction laws with high probability. Moreover, such estimators yield approximate systems, whose dynamics are approximations to the dynamics of the original system within the training time interval $[T_0, T]$, but can also provide approximations for emergent behaviors of collective dynamics, ranging from first-order opinion dynamics to second-order gravitational dynamics governing

the planetary movement in our solar system. A key component of evaluating the emergent dynamics are appropriate measures of the presence of a specific emergent behavior, which will be discussed in 3.4.

3.3 Learning Algorithm

Similar to the the algorithm presented in [89], the learning algorithm which we use for the more complex dynamics considered here starts from the introduction of suitable cost functions whose minimizers, over a suitable approximation space, determine the estimators. Equation (3.2.1) can be rewritten in a more compact form:

$$\begin{cases} \dot{\boldsymbol{X}} &= \mathbf{f}^{\text{nc},\boldsymbol{x}}(\boldsymbol{X},\boldsymbol{\Xi}) + \mathbf{f}^{\boldsymbol{\phi}^E}(\boldsymbol{X},\boldsymbol{\Xi}) \\ \dot{\boldsymbol{\Xi}} &= \mathbf{f}^{\text{nc},\xi}(\boldsymbol{X},\boldsymbol{\Xi}) + \mathbf{f}^{\boldsymbol{\phi}^\xi}(\boldsymbol{X},\boldsymbol{\Xi}). \end{cases}$$

Here $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T & \cdots & \boldsymbol{x}_N^T \end{bmatrix}^T \in \mathbb{R}^{Nd}$, $\boldsymbol{\Xi} = \begin{bmatrix} \xi_1 & \cdots \xi_N \end{bmatrix}^T \in \mathbb{R}^N$; for the interaction kernels, we use the vectorized notations, $\boldsymbol{\phi}^E = \{\phi^E_{k,k'} \in \mathcal{H}^{\boldsymbol{x}}_{k,k'}\}^K_{k,k'=1}$ and $\boldsymbol{\phi}^\xi = \{\phi^\xi_{k,k'} \in \mathcal{H}^\xi_{k,k'}\}^K_{k,k'=1}$, and $\mathbf{f}^{\boldsymbol{\phi}^E}, \mathbf{f}^{\boldsymbol{\phi}^\xi}$ are the collection of the corresponding right hand side terms in (3.2.1) respectively. Lastly, $\mathbf{f}^{\text{nc},\boldsymbol{x}}(\boldsymbol{X},\boldsymbol{\Xi})$ is defined as the vectorization of the non-collective forces $\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i,\xi_i) \in \mathbb{R}^d$ and $\mathbf{f}^{\text{nc},\xi}(\boldsymbol{X},\boldsymbol{\Xi})$ is defined as the vectorization of the non-collective forces $\mathbf{F}^\xi(\boldsymbol{x}_i,\xi_i) \in \mathbb{R}$. Our estimators are defined as the minimizers of the loss functions

$$\begin{cases} \widehat{\boldsymbol{\phi}}^E &= \underset{\boldsymbol{\varphi}^E \in \mathcal{H}^{\boldsymbol{x}}}{\arg\min} \sum_{m,l=1}^{M,L} \frac{1}{LM} ||\dot{\boldsymbol{X}}^m(t_l) - \mathbf{f}^{\text{nc},\boldsymbol{x}}(\boldsymbol{X}^m(t_l),\boldsymbol{\Xi}^m(t_l)) - \mathbf{f}^{\boldsymbol{\varphi}^E}(\boldsymbol{X}^m(t_l),\boldsymbol{\Xi}^m(t_l))||^2_{\mathcal{S}(d)} \\ \widehat{\boldsymbol{\phi}}^\xi &= \underset{\boldsymbol{\varphi}^\xi \in \mathcal{H}^\xi}{\arg\min} \sum_{m,l=1}^{M,L} \frac{1}{LM} ||\dot{\boldsymbol{\Xi}}^m(t_l) - \mathbf{f}^{\text{nc},\xi}(\boldsymbol{X}^m(t_l),\boldsymbol{\Xi}^m(t_l)) - \mathbf{f}^{\boldsymbol{\varphi}^\xi}(\boldsymbol{X}^m(t_l),\boldsymbol{\Xi}^m(t_l))||^2_{\mathcal{S}(1)} \end{cases},$$

where the $\|\cdot\|_{\mathcal{S}(\cdot)}$ norm is defined as

$$\|\boldsymbol{Z}\|^2_{\mathcal{S}(d')} = \sum_{i=1}^{N} \frac{1}{N_{k_i}} \|\boldsymbol{z}_i\|$$

for $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^T & \dots \boldsymbol{z}_N^T \end{bmatrix}^T$ with each $\boldsymbol{z}_i \in \mathbb{R}^{d'}$ ($d' = d$ or 1). Here $\|\cdot\|$ is the same norm used in (3.2.1) and (3.2.2); $\mathcal{H}^{\boldsymbol{x}} = \bigoplus_{k,k'=1}^{K} \mathcal{H}^{\boldsymbol{x}}_{k,k'}$ and $\mathcal{H}^{\xi} = \bigoplus_{k,k'=1}^{K} \mathcal{H}^{\xi}_{k,k'}$ are finite-dimensional hypothesis spaces. We choose each of the hypothesis space $\mathcal{H}^{\boldsymbol{x}}_{k,k'}$ be to a finite dimensional function space of piece-wise polynomials of degree p, with $p = 0$ or 1 (polynomials of higher degree can be used and other type of basis functions are also possible, e.g., clamped B-splines, see [89]), with polynomial pieces supported on intervals that form a uniform partition of the observed range of variables $[R^{\boldsymbol{x}}_{k,k',\min}, R^{\boldsymbol{x}}_{k,k',\max}] \times [S^{\boldsymbol{x}}_{k,k',\min}, S^{\boldsymbol{x}}_{k,k',\max}]$. Hence, each $\varphi^E_{k,k'}$ can be expressed in terms of the linear combination of the basis functions as follows

$$\varphi^E_{k,k'}(r, s^{\boldsymbol{x}}) = \sum_{\eta^{\boldsymbol{x}}_{k,k'}=1}^{n^{\boldsymbol{x}}_{k,k'}} \alpha^{\boldsymbol{x}}_{k,k',\eta^{\boldsymbol{x}}_{k,k'}} \psi^{\boldsymbol{x}}_{k,k',\eta^{\boldsymbol{x}}_{k,k'}}(r, s^{\boldsymbol{x}}).$$

Similar definitions are used for each $\mathcal{H}^{\xi}_{k,k'}$. Substituting this expression into the functionals above, the minimization becomes a set of linear equations,

$$A^{\boldsymbol{x}} \vec{\alpha}^{\boldsymbol{x}} = \vec{b}^{\boldsymbol{x}} \quad \text{and} \quad A^{\xi} \vec{\alpha}^{\xi} = \vec{b}^{\xi}.$$

Here $\vec{\alpha}^{\boldsymbol{x}} \in \mathbb{R}^{n^{\boldsymbol{x}}}$ is the vector of $\alpha^{\boldsymbol{x}}_{k,k',\eta^{\boldsymbol{x}}_{k,k'}}$'s, and $A^{\boldsymbol{x}} \in \mathbb{R}^{n^{\boldsymbol{x}} \times n^{\boldsymbol{x}}}$ with $n^{\boldsymbol{x}} = \sum_{k,k'=1}^{K} \eta^{\boldsymbol{x}}_{k,k'}$; similarly for $\vec{\alpha}^{\xi}$ and A^{ξ}.

Remark 3.3.1. *In the case of missing $\dot{\boldsymbol{x}}_i(t)$ (for first order system) or $\ddot{\boldsymbol{x}}_i(t)$ (for second order system), we will approximate it using an appropriate finite difference scheme. See Sec. 3.4.4 for details on how we setup the examples with or without*

derivative information.

In the case of the second-order dynamics described in (3.2.2), we introduce a new variable $\boldsymbol{v}_i(t) = \dot{\boldsymbol{x}}_i(t) \in \mathbb{R}^d$ and let $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1^T & \dots & \boldsymbol{v}_N^T \end{bmatrix}^T$, a compact form of (3.2.2) is given as follows,

$$\begin{cases} \dot{\boldsymbol{X}} &= \boldsymbol{V} \\ \dot{\boldsymbol{V}} &= \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\Xi}) + \mathbf{f}^{\boldsymbol{\phi}^E}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\Xi}) + \mathbf{f}^{\boldsymbol{\phi}^A}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\Xi}) \in \mathbb{R}^{Nd} \\ \dot{\boldsymbol{\Xi}} &= \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\Xi}) + \mathbf{f}^{\boldsymbol{\phi}^\xi}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\Xi}) \in \mathbb{R}^{N} \end{cases}$$

Here $\boldsymbol{\phi}^A = \{\phi^A_{k,k'} \in \mathcal{H}^{\dot{\boldsymbol{x}}}_{k,k'}\}_{k,k'=1}^K$. We find the estimators from the following minimizations

$$\begin{cases} (\widehat{\boldsymbol{\phi}}^E, \widehat{\boldsymbol{\phi}}^A) &= \underset{\boldsymbol{\varphi}^E \in \mathcal{H}^{\boldsymbol{x}}, \boldsymbol{\varphi}^A \in \mathcal{H}^{\dot{\boldsymbol{x}}}}{\arg\min} \{\sum_{m,l=1}^{M,L} \frac{1}{LM} ||\dot{\boldsymbol{V}}^m(t_l) - \mathbf{f}^{\mathrm{nc},\dot{\boldsymbol{x}}}(\boldsymbol{X}^m(t_l),\boldsymbol{V}^m(t_l),\boldsymbol{\Xi}^m(t_l)) \\ & \qquad\qquad -\mathbf{f}^{\boldsymbol{\varphi}^E}(\boldsymbol{X}^m(t_l),\boldsymbol{V}^m(t_l),\boldsymbol{\Xi}^m(t_l)) \\ & \qquad\qquad -\mathbf{f}^{\boldsymbol{\varphi}^A}(\boldsymbol{X}^m(t_l),\boldsymbol{V}^m(t_l),\boldsymbol{\Xi}^m(t_l))||^2_{\mathcal{S}(d)}\} \\ \widehat{\boldsymbol{\phi}}^\xi &= \underset{\boldsymbol{\varphi}^\xi \in \mathcal{H}^\xi}{\arg\min} \{\sum_{m,l=1}^{M,L} \frac{1}{LM} ||\dot{\boldsymbol{\Xi}}^m(t_l) - \mathbf{f}^{\mathrm{nc},\xi}(\boldsymbol{X}^m(t_l),\boldsymbol{V}^m(t_l),\boldsymbol{\Xi}^m(t_l)) \\ & \qquad\qquad -\mathbf{f}^{\boldsymbol{\varphi}^\xi}(\boldsymbol{X}^m(t_l),\boldsymbol{V}^m(t_l),\boldsymbol{\Xi}^m(t_l))||^2_{\mathcal{S}(1)}\} \end{cases}$$

Here $\mathcal{H}^{\dot{\boldsymbol{x}}} = \bigoplus_{k,k'=1}^K \mathcal{H}^{\dot{\boldsymbol{x}}}_{k,k'}$. By choosing appropriate finite dimensional hypothesis spaces for $\mathcal{H}^{\boldsymbol{x}}, \mathcal{H}^{\dot{\boldsymbol{x}}}$ and $\mathcal{H}^\xi$, e.g., piece-wise polynomials of degree p, we can simplify the least square problems down to the following linear systems which we solve to generate the necessary coefficients:

$$A\vec{\alpha} = \vec{b} \quad \text{and} \quad A^\xi \vec{\alpha}^\xi = \vec{b}^\xi.$$

Here, $\vec{\alpha} = \begin{bmatrix} (\vec{\alpha}^{\boldsymbol{x}})^T & (\vec{\alpha}^{\dot{\boldsymbol{x}}})^T \end{bmatrix}^T$ with $\vec{\alpha}^{\boldsymbol{x}}$ being the collection of $\alpha^{\boldsymbol{x}}_{k,k',\eta^{\boldsymbol{x}}_{k,k'}}$'s and $\vec{\alpha}^{\dot{\boldsymbol{x}}}$ being

the collection of $\alpha^{\dot{\boldsymbol{x}}}_{k,k',\eta^{\dot{\boldsymbol{x}}}_{k,k'}}$'s.

Remark 3.3.2. *For further details regarding the construction of the learning matrices, A, A^{ξ}, and the right hand side vectors, $\vec{b}, \vec{b}^{\xi}$, we refer the reader to Section 2 : Algorithm of the Supplementary Information of [89].*

3.3.1 Computational Complexity

The learning approach, which is described in Sec. 3.3, can be easily parallelized in the m (number of initial conditions) variable. Although it takes MLD double-precision floating-point numbers ($D = Nd + N$ for a first-order system, and $D = 2Nd + N$ for a second-order system) to store the discrete trajectory data, each computing core j only needs to store $M_j LD$ floating-point numbers, with $M_j \approx \frac{M}{\text{Number of Cores}}$. Furthermore, each computing core does not need to hold all of the trajectory data in memory, since the assembly of the learning matrix and the right-hand-side vector needs only LD floating-point numbers (one system trajectory at a time). The sizes for the learning matrix and right hand side vector are: $n \times n$ and $n \times 1$ ($n = n^{\boldsymbol{x}}$ or $n = n^{\xi}$ for a first-order system and $n = n^{\boldsymbol{x}} + n^{\dot{\boldsymbol{x}}}$ or $n = n^{\xi}$ for a second-order system), respectively. Since we have $n^2 \ll LD$, $n^2 \ll MLD$, which makes solving for our estimators extremely memory efficient. At each time instance, we have to compute the various pairwise variables, requiring $\mathcal{O}(N^2)$ distance calculations, hence the algorithm performs a total of $\mathcal{O}(MLN^2)$ computations of pairwise variables. In solving the linear system, it performs $\mathcal{O}(n^3)$ operations (or $\mathcal{O}(n^2 \log(n))$, we take the worst cases scenario since we use the built-in pseudo-inverse routine in MATLAB to avoid any possible issues with numerical stability). The total computational complexity is $\mathcal{O}(MLN^2 + n^3)$. Online learning can be built into our learning approach: as trajectory data from different initial conditions comes in, one can simply average the estimators from previous trajectory data with the estimators from the new trajectory data to obtain a better approximation.

3.4 Performance Measures

We consider three different kinds of performance measures: how close the estimated interaction kernel(s) are to the true one(s), how well the trajectories of the system driven by the estimated interaction kernel(s) approximate the trajectories of the original system, and finally how well emergent patterns are reproduced/predicted in the system driven by the estimated interaction kernels. We use appropriate dynamics adapted measures, specified in section 3.4.1 and the Appendix.

3.4.1 Estimation error of interaction kernels

Following the definitions in [89], we introduce a set of probability measures to calculate the learning error between $\widehat{\boldsymbol{\phi}}^E$ and $\hat{\phi}^E$, for any first-order system. We define the following probability measures, $\rho_T^{E,k,k'}, \rho_T^{L,E,k,k'}, \rho_T^{L,M,E,k,k'}$, to measure the performance of our estimators. The first-order measures are as follows,

$$\begin{cases} \rho_T^{E,k,k'}(r, s^{\boldsymbol{x}}) & = \frac{1}{N_{k,k'}T} \int_{t=0}^{T} \mathbb{E}_{\mathbf{Y}_0 \sim \mu^{\boldsymbol{y}}} \Big[\sum_{\substack{i \in C_k \\ i' \in C_{k'} \\ i \neq i'}} \delta_{r_{i,i'}(t), s^{\boldsymbol{x}}_{i,i'}(t)}(r, s^{\boldsymbol{x}}) \Big] \, dt, \\ \rho_T^{L,E,k,k'}(r, s^{\boldsymbol{x}}) & = \frac{1}{N_{k,k'}L} \sum_{l=1}^{L} \mathbb{E}_{\mathbf{Y}_0 \sim \mu^{\boldsymbol{y}}} \Big[\sum_{\substack{i \in C_k \\ i' \in C_{k'} \\ i \neq i'}} \delta_{r_{i,i'}(t_l), s^{\boldsymbol{x}}_{i,i'}(t_l)}(r, s^{\boldsymbol{x}}) \Big], \\ \rho_T^{L,M,E,k,k'}(r, s^{\boldsymbol{x}}) & = \frac{1}{N_{k,k'}LM} \sum_{l,m=1}^{L,M} \sum_{\substack{i \in C_k \\ i' \in C_{k'} \\ i \neq i'}} \delta_{r_{i,i'}(t_l), s^{\boldsymbol{x}}_{i,i'}(t_l)}(r, s^{\boldsymbol{x}}). \end{cases} \tag{3.4.1}$$

$\rho_T^{E,k,k'}$ (for continuous trajectory) and $\rho_T^{L,E,k,k'}$ (for discrete trajectory) are only used in the theoretical setting; in practice, we use $\rho_T^{L,M,E,k,k'}$ (with large M and L) for actual implementations and applications. These measures depend on the dynamical system and the distribution of initial conditions, weighting the areas of pairwise distances (the variable r) and of variables $s^{\boldsymbol{x}}$ based on how often trajectories of the system explore them.

Table 3.3 explains the definitions of the variables in (3.4.1).

Variable	Definition
$\boldsymbol{Y}$	$\begin{bmatrix}\boldsymbol{X}^T & \boldsymbol{\Xi}^T\end{bmatrix}^T$
$\mu^{\boldsymbol{y}}$	$\begin{bmatrix}\mu^{\boldsymbol{x}} & \mu^{\xi}\end{bmatrix}^T$
$r_{i,i'}(t)$	$\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|$
$s^{\boldsymbol{x}}_{i,i'}(t)$	$\mathcal{F}^{\boldsymbol{x}}(\boldsymbol{x}_i(t), \xi_i(t), \boldsymbol{x}_{i'}(t), \xi_{i'}(t)) : \mathbb{R}^{2d+2} \to \mathbb{R}$
$N_{k,k'}$	$\begin{cases} N_k(N_k-1) & \text{if } k = k', \\ N_k N_{k'} & \text{if } k \neq k'. \end{cases}$

Table 3.3: ρ_T's, Definition of the Variables

In the case of $N_{k,k'} = 0$, we define the corresponding $\rho_T^{E,k,k'}(r, s^{\boldsymbol{x}})$ to a zero function.

We measure the error of the interaction kernel estimators, $\phi^E_{k,k'} - \hat{\phi}^E_{k,k'}$, using the dynamics-induced weighted L^2 norm

$$\left\|\phi^E_{k,k'} - \hat{\phi}^E_{k,k'}\right\|^2_{L^2(\rho_T^{E,k,k'})} = \int_{r=0}^{\infty}\int_{s^{\boldsymbol{x}}=-\infty}^{\infty} (\phi^E_{k,k'}(r, s^{\boldsymbol{x}}) - \hat{\phi}^E_{k,k'}(r, s^{\boldsymbol{x}})^2\, r^2 d\rho_T^{E,k,k'}(r, s^{\boldsymbol{x}}). \tag{3.4.2}$$

However since $\rho_T^{E,k,k'}$ is not calculable, we use $\rho_T^{L,M,E,k,k'}$ instead. The weight, r^2, comes from the governing structure of (3.2.1). Theoretical guarantees such as those in [89, 90] bound these errors, with high probability, as M grows. Extending those bounds to the general types of systems considered here will be investigated in future work. The results of our numerical experiments suggest that the learning rate, i.e. the rate of decrease of the error in (3.4.2), as a function of M is independent of the dimension of the state space of the system, and only depends crucially on the number of variables in the interaction kernel. The curse of dimensionality (of the state space) is therefore avoided.

3.4.2 Trajectory errors

We consider another performance measure, which might be estimated from data, especially when the true interaction kernel is not known, that quantifies the prediction capability of our estimators, by comparing the observed trajectories to the estimated trajectories evolved from the same initial conditions but using the esti-

mated interaction laws. We will consider both $\boldsymbol{X}(t) = \begin{bmatrix} \boldsymbol{x}_1^T(t) & \cdots \boldsymbol{x}_N^T(t) \end{bmatrix}^T$ and $\boldsymbol{\Xi}(t) = \begin{bmatrix} \xi_1(t) & \cdots & \xi_N(t) \end{bmatrix}^T$ for $t \in [T_0, T]$. Let $\boldsymbol{X}_{[T_0,T]} = \{\boldsymbol{X}(t)\}_{t\in[T_0,T]}$, then the following norm is used

$$\left\| \boldsymbol{X}_{[T_0,T]} - \hat{\boldsymbol{X}}_{[T_0,T]} \right\|_{\mathcal{T}(d)} = \frac{\max_{t\in[0,T]} \left\| \boldsymbol{X}(t) - \hat{\boldsymbol{X}}(t) \right\|_{\mathcal{S}(d)}}{\max_{t\in[0,T]} \left\| \boldsymbol{X}(t) \right\|_{\mathcal{S}(d)}}. \tag{3.4.3}$$

Here $\hat{\boldsymbol{X}}_{[T_0,T]}$ is the estimated trajectory using our estimators with the same initial condition as in $\boldsymbol{X}_{[T_0,T]}$. The scaling by $\max_{t\in[0,T]} \left\| \boldsymbol{X}(t) \right\|_{\mathcal{S}(d)}$ enables us to compare trajectory errors for different kinds of dynamics, especially those with large $\|\boldsymbol{x}_i\|$. Similarly,

$$\left\| \boldsymbol{V}_{[T_0,T]} - \hat{\boldsymbol{V}}_{[T_0,T]} \right\|_{\mathcal{T}(1)} = \frac{\max_{t\in[0,T]} \left\| \boldsymbol{V}(t) - \hat{\boldsymbol{V}}(t) \right\|_{\mathcal{S}(d)}}{\max_{t\in[0,T]} \left\| \boldsymbol{V}(t) \right\|_{\mathcal{S}(d)}}, \tag{3.4.4}$$

and

$$\left\| \boldsymbol{\Xi}_{[T_0,T]} - \hat{\boldsymbol{\Xi}}_{[T_0,T]} \right\|_{\mathcal{T}(1)} = \frac{\max_{t\in[0,T]} \left\| \boldsymbol{\Xi}(t) - \hat{\boldsymbol{\Xi}}(t) \right\|_{\mathcal{S}(1)}}{\max_{t\in[0,T]} \left\| \boldsymbol{\Xi}(t) \right\|_{\mathcal{S}(1)}}. \tag{3.4.5}$$

For performance measures defined for $\widehat{\boldsymbol{\phi}}^{\xi}$ and the second order systems, please see sec. 3.9 in the appendix.

3.4.3 Confusion Matrix and Pattern Indicator Scores

When a system is highly sensitive on small perturbations, or even chaotic, it is hopeless to expect that the estimated system will produce trajectories that are accurate approximations of the trajectories of the original system, except perhaps for very small times. However we have observed that certain large-time aspects of the dynamics of the system, such as certain emergent behavior including flocking or milling or clustering, are preserved in our estimated system, even when the trajectory-wise errors are relatively large. On the one hand, this may seem surprising, as at no point do we inject any knowledge about such emergent behaviors, into our estimator; on the

other hand if such emergent behaviors are thought of as being "structurally robust" to perturbations of the system, and even perturbations of the laws of the system (the interaction kernels), then it becomes reasonable to expect that our estimated systems should preserve, at least to some degree, such emergent behaviors. We therefore introduce a way of measuring quantitatively the presence of such emergent behaviors, and quantify the performance in reproducing them in our estimated systems.

In order to accurately describe the capability of our estimators to predict the correct *emergent* behaviors at large time $T_f \gg T$, we consider confusion matrices and "pattern indicator scores". These are defined differently for each dynamical system to measure its unique emergent behavior.

Several aspects of the emergent behaviors that we are interested in are observables (i.e. functions defined on the state variables of the system). We define various emergent behavior scores, such as the flocking score, the milling score, etc., and choose a target range for the score to be in as an indicator of occurrence of the emergent behavior. For example, if the flocking score is within $(0.99, 1]$, then flocking occurs. We calculate these scores on the true and estimated systems (systems with the same initial conditions as the true systems but evolved using the learned interaction law(s)). From this indicator of whether the emergent behavior occurred in the true/estimated system, we construct a confusion matrix, given as follows (in the case of learning flocking systems),

	Predicted Non-flocking	Predicted flocking
True Non-flocking	$p_{1,1}$	$p_{1,2}$
True flocking	$p_{2,1}$	$p_{2,2}$

Table 3.4: General form of a confusion matrix. Each $p_{i,j}$ shows a probability (represented in percentages) of the combination, e.g., $p_{1,1}$ is the probability of the predicted system (evolved using the estimated interaction laws) showing non-flocking behavior given that the true system shows non-flocking behavior with the same initial conditions.

It is used to present the probability of the occurrence of the desired emergent behaviors in the true and estimated systems. Namely, if the true systems exhibit

flocking with high probability, then the estimated systems should ideally show flocking with similar probability.

In order to provide deeper insight about the prediction of emergent behavior via confusion matrices, we also consider the following statistics from the confusion matrix.

Accuracy	Precision	Recall	F Score
$\frac{p_{1,1}+p_{2,2}}{\sum_{1=i,j=2} p_{i,j}}$	$\frac{p_{2,2}}{p_{2,1}+p_{2,2}}$	$\frac{p_{2,2}}{p_{1,2}+p_{2,2}}$	$\frac{2}{\frac{1}{\text{Prec.}}+\frac{1}{\text{Recal.}}}$

Table 3.5: Definition of the statistical terms related to the confusion matrix.

Next we use a more refined measurement, a so-called 'pattern indicator score', to further demonstrate the capabilities of the estimated system at predicting emergent behaviors. Besides the emergent behavior scores, there are other quantitative descriptions of the emergent behaviors, such as the center-of-mass velocity in flocking, the common rotational axis in milling, the conservation of total energy in concentric trajectories, etc. The pattern indicator scores use these, sometimes together with the previously defined emergent behavior scores, to measure how well the estimated systems are predicting these observables compared to the true systems. Details of the definition of the confusion matrices and pattern indicator scores for each dynamics are in Sec. 3.5 to Sec. 3.7.

3.4.4 Setup of the Numerical Experiments

Here we describe the general setup for the subsequent sections of experiments. The various dynamical systems we consider exhibit a wide variety of emergent behaviors: clustering, flocking, milling, synchronization, and concentric trajectories. Different forms of interaction kernels are also considered, i.e., $\phi(r)$, $\phi(r,s)$ and $\phi(r;\boldsymbol{P})$, where $\boldsymbol{P}$ is an unknown vector of parameters. These dynamics range from first-order dynamics of homogeneous agents to second-order dynamics of heterogeneous agents. We arrange the examples in three major sections based on the different types of the interaction laws.

The experiments are setup as follows: we first run M_ρ different initial conditions generated i.i.d from the probability measure $\mu^{\boldsymbol{y}}$ for initial condition, and evolve[1] the dynamics from 0 to T: the dynamics observed in $[T_0, T]$ is used to compute the probability measures ρ_T^L's, which are empirical approximations to the probability measures ρ_T's. We do this only to compute and report the $L^2(\rho_T)$ approximation errors; in practice this step is not required nor needed. Next, we generate another set of M random initial conditions and corresponding trajectories of the dynamics for $t \in [0, T]$, with each dynamics observed at L equidistant times $T_0 = t_1 < t_1 < \cdots < t_L = T$, producing the observation data, i.e. $\{\boldsymbol{y}_i(t_l)^m\}_{i,l,m=1}^{N,L,M}$, without the corresponding derivative information (i.e., $\dot{\boldsymbol{y}}_i(t_l)^m$ is not given, except for Synchronized Oscillator Dynamics and Gravitational Dynamics), as input to our estimation procedure. We construct the hypothesis spaces, where the estimators are found, on the learning intervals, e.g. $[r_{\min}^{k,k'}, r_{\max}^{k,k'}] \times [s_{\min}^{k,k'}, s_{\max}^{k,k'}]$, derived from the observation data; the numbers of basis functions, as well as their degrees, are reported in each section. We report the $L^2(\rho_T)$ errors between the estimated and true interaction kernels, as well as the trajectory errors based on the statistics over the training set and over a testing set (with new initial conditions), in the form of (mean value)$\pm$(standard deviation). Then we consider the emergent behavior of the true dynamics and the predicted dynamics at $T_f \gg T$, and evaluate "pattern indicator scores" and confusion matrices corresponding to the various kinds of emergent behaviors. The parameters used by all experiments are reported in table 3.6.

N	M_ρ	T_0	L	# Learning Trials
20	2000	0	500	10

Table 3.6: Common Parameters

Each section/subsection is presented in a similar manner: we introduce the model

[1]The evolution of the dynamical system is done using the built-in integrator, ode15s, of MATLAB with the relative tolerance set at 10^{-8} and absolute tolerance set at 10^{-11}.

and discuss why such model interesting for our learning approach; then, we relate the model equation to the learning paradigm presented in (3.2.1) and (3.2.2). Next, we present our learning results in figures and tables, in terms of approximation error, trajectory error, confusion matrix and pattern indicators. We end with a brief discussion of the learning results.

3.5 Emergent Behaviors Induced by $\phi(r)$

We consider here three prototypical types of emergent behavior: clustering, flocking and milling. We examined four different examples of collective dynamics in order to comprehensively explore all three types of emergent behaviors, with very different dynamical behaviors.

3.5.1 Opinion Dynamics

The opinion dynamics (OD) model, first introduced in [76], is a prototypical first-order model of homogeneous agents which describes the interaction of people's opinions through time, see details and extensions in [40, 16, 98, 20, 71, 57]. These models have gained popularity in modeling human's social behavior, and they can be used to predict interesting social phenomena, namely, clustering/consensus of opinions.

The governing equations ($\boldsymbol{x}_i \in \mathbb{R}^d$ being a vector of opinions) are:

$$\dot{\boldsymbol{x}}_i = \sum_{i'=1}^{N} \frac{1}{N} \phi^E(\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|)(\boldsymbol{x}_{i'} - \boldsymbol{x}_i), \quad \text{for } i = 1, \cdots, N.$$

Here $\phi^E(r) \geq 0$ for all $r \geq 0$. With the interaction kernels giving attractive influences only, these models are bound to have clusters of opinions at large time. Table 3.7a shows how this dynamical system is mapped to the general form (3.2.1). The parameters used for setting up the experiment used are shown in table 3.7b.

Category	ξ_i	K	$s^{\boldsymbol{x}}_{i,i'}$	$\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i)$	ϕ^E
Value	$\emptyset$	1	$\emptyset$	$\emptyset$	non-negative

(a) (OD) Mapping to (3.2.1)

M	d	T_f	T	$\mu^{\boldsymbol{x}}$
250	2	50	10	Unif. on $[0,5]^2$

(b) (OD) Parameters for Experiment Setup

Table 3.7: Opinion Dynamics (OD)

We consider the following interaction law,

$$\phi^E(r) = \begin{cases} 1 & \text{if } 0 \le r < \frac{1}{\sqrt{2}} \\ 0.1 & \text{if } \frac{1}{\sqrt{2}} \le r < 1 \\ 0 & \text{otherwise} \end{cases}$$

Piece-wise constant polynomials with $n^{\boldsymbol{x}} = 99$ basis functions are used to approximate ϕ^E. The comparison of the true ϕ^E and the estimated $\hat{\phi}^E$ is shown in Fig.3.1.

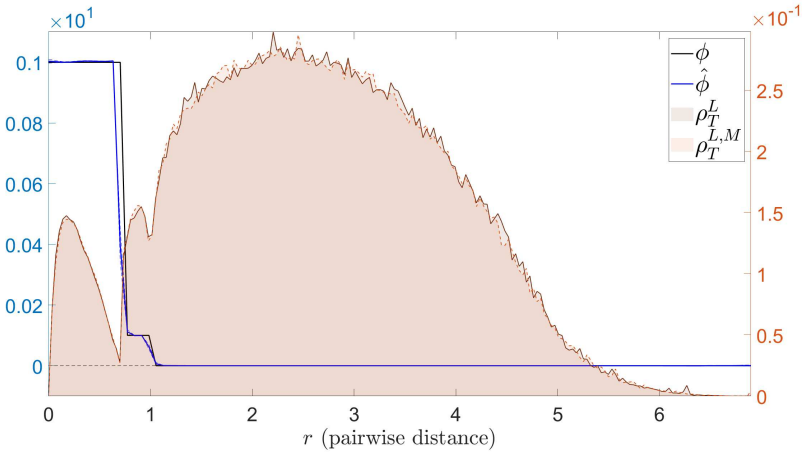


Figure 3.1: (OD) Comparison of ϕ^E and $\hat{\phi}^E$, with the relative error being $1.4\cdot10^{-1}\pm2\cdot10^{-2}$ (calculated using (3.4.2)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in blue dotted lines. Shown in the background is the comparison of approximated ρ_T^L versus the empirical $\rho_T^{L,M}$.

As it is shown in Fig. 3.1, not only can our estimator detect the discontinuity in the ϕ, but also can it detect the compact support of ϕ. Meanwhile, there is higher uncertainty in learning the interaction kernel at $r = 0$ (the information of $\phi^E(0)$ is lost since it is weighted by corresponding $\boldsymbol{r}_{i,i'}$) and at those discontinuity points. Since ϕ^E

is non-negative, the agents in the system would eventually converge to clusters, this decreases the effective number of pairwise distance data for inferring ϕ^E. However, we are still able to provide an accurate estimator of ϕ^E by the continuity of the estimator. The comparison of a trajectory driven by the true ϕ^E versus the other one driven by the estimated $\hat{\phi}^E$ is shown in Fig. 3.2: there is no major visual difference between the true and predicted trajectories (generated from the training initial condition); the differences are quantified in table 3.8.

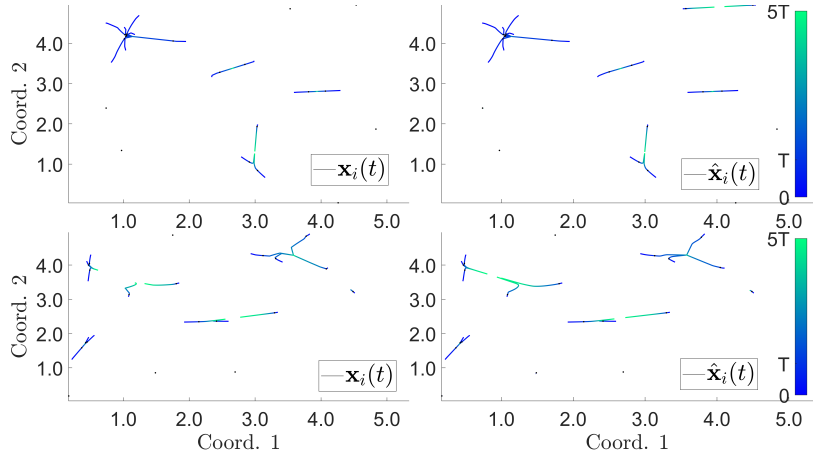


Figure 3.2: (OD) Comparison of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, with the errors reported in table 3.8. The first row of trajectories are generated from an initial condition taken from the observation data. The second row of trajectories are generated from another randomly chosen initial condition. The first column of trajectories are generated from the true interaction kernel, whereas the second column of trajectories are generated from our estimated kernel with the same initial conditions. The color of the trajectory indicates the flow of time, from deep blue (at $t = T_0$) to light green (at $t = T_f$).

	$[0, T]$	$[T, T_f]$
mean_{IC}: Training ICs	$6.5 \cdot 10^{-3} \pm 7 \cdot 10^{-4}$	$2.5 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
std_{IC}: Training ICs	$7.4 \cdot 10^{-3} \pm 7 \cdot 10^{-4}$	$2.5 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
mean_{IC}: Random ICs	$6.6 \cdot 10^{-3} \pm 6 \cdot 10^{-4}$	$2.73 \cdot 10^{-2} \pm 9 \cdot 10^{-4}$
std_{IC}: Random ICs	$7.4 \cdot 10^{-3} \pm 1 \cdot 10^{-3}$	$2.7 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$

Table 3.8: (OD) Trajectory Errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory errors is calculated using (3.4.3).

The confusion matrix and pattern indicator scores used to examine the capability of our estimators predicting the proper emergent behaviors associated with the Opinion

Dynamics model are defined as follows. First, a confusion matrix is used to show the accuracy of our estimator to display the same clustering behavior as the true systems, see the results in table 3.9.

	Predicted Non-Clustering	Predicted Clustering
True Non-Clustering: Training ICs	$88 \pm 2\%$	$2 \pm 1\%$
True Clustering: Training ICs	$1.2 \pm 0.6\%$	$8.9 \pm 0.2\%$
True Non-Clustering: Random ICs	$88 \pm 2\%$	$1.6 \pm 0.9\%$
True Clustering: Random ICs	$1.7 \pm 0.8\%$	$8 \pm 2\%$

Table 3.9: (OD) Confusion Matrix: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). It is generated using table 3.4.

We provide more statistics about the confusion matrix in order to understand our prediction of clustering better in table 3.10.

	Accuracy	Precision	Recall	F-Score
Training ICs	$97 \pm 1.5\%$	$84 \pm 11.0\%$	$88.0 \pm 4.9\%$	$85.5 \pm 7.1\%$
Random ICs	$96.6 \pm 1.2\%$	$83.6 \pm 7.5\%$	$82.5 \pm 8.4\%$	$82.8 \pm 6.5\%$

Table 3.10: (OD) Confusion Matrix Statistics: ICs used in the training set, new ICs randomly drawn from $\mu^{\boldsymbol{x}}$. It is generated using table 3.5.

Next, when the true system has clustering, we want to know if the predicted system can have the same number of clusters as the true system has. Hence, we assign a score of 1 when the predicted system shows the same number of clusters as the true systems; and a score of 0 when it predicts the wrong number of clusters. PI_1 is the average of those scores over M trials. Lastly, we want to compare the clusters between the true and predicted systems. Let $\mathcal{C}$ contain the centers of the clusters at time T[2] from the true system, $\hat{\mathcal{C}}$ contain the centers of clusters from the estimated system; we shall use Hausdorff distance to calculate the distance between $\mathcal{C}$ and $\hat{\mathcal{C}}$. PI_2 is the average of M trials of such distances. See table 3.11 for details.

[2]The clusters are collection of points such that for each cluster $\mathcal{C}$, if $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{C}$, then $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| < \delta$, and if $\boldsymbol{x}_i \in \mathcal{C}$ and $\boldsymbol{x}_j \notin \mathcal{C}$, then $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| > \delta$. Here we chose $\delta = 0.01$.

	PI_1	PI_2
$mean_{IC}$: Training ICs	$9.2 \cdot 10^{-1} \pm 2 \cdot 10^{-2}$	$3.2 \cdot 10^{-2} \pm 3 \cdot 10^{-3}$
std_{IC}: Training ICs	$2.7 \cdot 10^{-1} \pm 3 \cdot 10^{-2}$	$4.3 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
$mean_{IC}$: Random ICs	$9.3 \cdot 10^{-1} \pm 2 \cdot 10^{-2}$	$3.4 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$
std_{IC}: Random ICs	$2.6 \cdot 10^{-1} \pm 3 \cdot 10^{-2}$	$4.5 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$

Table 3.11: (OD) Pattern Indicator Scores: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows).

The smaller PI_1 is, the better we are at predicting the number of clusters right. Meanwhile, not only could we predict the number of clusters with high confidence, but also could we predict the actual location of the clusters.

3.5.2 Cucker-Smale Dynamics

Modeling how animals (or other living agents) move in a cohesive group formation has been a challenging and well-studied problem [46, 47, 35], [38, 61, 37], [99, 74, 128]. There are different degress of cohesion in a collective system: flocking (where each agent shares a common velocity), milling (where each agent rotates around the same axis or about the same oint), and swarming (transition state between flocking and milling). We first consider the simplest cohesion in a collective system, namely flocking (see detailed work in [49, 48, 41, 42, 4, 35, 137], its mean field limit in [65, 123], and extension to a stochastic system in [59] and references therein), and investigate the learnability of these flocking systems.

The Cucker-Smale (CS) dynamics is one of the prototypical examples of flocking agents[3]. Its governing equations are

$$\ddot{\boldsymbol{x}}_i = \sum_{i'=1}^{N} a_{i,i'}(\boldsymbol{X})(\dot{\boldsymbol{x}}_{i'} - \dot{\boldsymbol{x}}_i), \quad \text{for } i = 1, \cdots N.$$

Here $a_{i,i'}(\boldsymbol{X}) = H \cdot (1 + \|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|^2)^{-\beta}$ where H, β are chosen parameters. Table 3.12

[3]The Vicsek model in [137] is a seminal work in modeling flocking of birds, but it uses a different paradigm (different from (3.2.2)) to model the flocking behavior.

shows how this dynamical system is mapped to the general form (3.2.2).

Category	m_i	ξ_i	K	$s^{\dot{\boldsymbol{x}}}_{i,i'}$	$\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i)$	ϕ^E	$\phi^A(r)$
Value	1	$\emptyset$	1	$\emptyset$	$\emptyset$	$\emptyset$	$\frac{H}{(1+r^2)^\beta}$

Table 3.12: (CS) Mapping to (3.2.2)

M	d	T_f	T	$\mu^{\boldsymbol{x}}$	$\mu^{\dot{\boldsymbol{x}}}$
500	2	50	5	Unif. on $[-5,5]^2$	Unif. on $[-5,5]^2$

Table 3.13: (CS) Parameters for Experiment Setup

With certain choices of H and β, the Cucker-Smale system is guaranteed to produce flocking (where all agents have the same final velocity) see [49]. For example, when $\beta < \frac{1}{2}$, the system is guaranteed to have flocking regardless of initial conditions; when $\beta = \frac{1}{2}$, the system has conditional flocking depending on the initial configuration of velocities; when $\beta > \frac{1}{2}$, the system has conditional flocking depending on the initial configuration of both positions and velocities.

We consider the following interaction law,

$$\phi^A(r) = \frac{1}{(1+r^2)^{\frac{1}{4}}}.$$

With this interaction kernel, the agents are guaranteed to flock (see theorem $2, 3$ in [49]). We use the following parameters in table 3.13 to set up the experiment. Piece-wise linear polynomials with $n^{\dot{\boldsymbol{x}}} = 100$ basis functions are used to approximate ϕ^A. The comparison of the true ϕ^A and the estimated $\hat{\phi}^A$ is shown in Fig.3.3. As it is shown in Fig. 3.3, our learning approach produces faithful approximation to ϕ^A, especially capturing the tail behavior of the original interaction law, notwithstanding the scarcity of samples in that region of pairwise distances and speeds; towards $r = 0$, the estimated kernel is also close to the true kernel. The comparison of true trajectory $\boldsymbol{X}(t)$ and learned $\hat{\boldsymbol{X}}(t)$ is shown in Fig. 3.4. Fig. 3.4 shows no visual difference

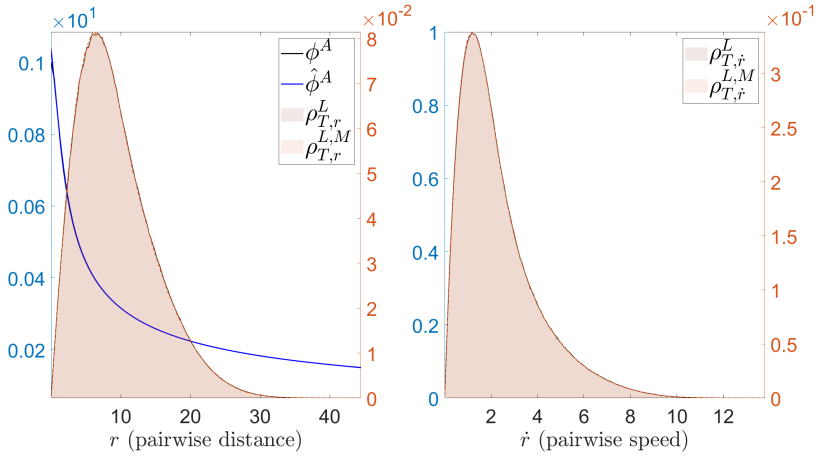


Figure 3.3: (CS) Comparison of ϕ^A and $\hat{\phi}^A$ together with a plot of $\rho^L_{T,\dot{r}}$ versus $\rho^{L,M}_{T,\dot{r}}$, with the relative error being $4.7 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$ (calculated using (3.9.6)). The true interaction kernel is shown by in a black solid line, whereas the mean estimated interaction kernel is shown in a blue solid line with its confidence interval shown in blue dotted lines. Shown in the background is the comparison of approximated ρ^L_T versus the empirical $\rho^{L,M}_T$.

between the true trajectories and the learned trajectories (for the training initial condition and a randomly chosen initial condition), we provide a quantitative insight into the difference between trajectories in table 3.14.

	$[0,T]$	$[T,T_f]$
mean_{IC}: Training ICs on $\boldsymbol{x}$	$1.55 \cdot 10^{-3} \pm 9 \cdot 10^{-5}$	$1.9 \cdot 10^{-3} \pm 1 \cdot 10^{-4}$
mean_{IC}: Training ICs on $\boldsymbol{v}$	$2.7 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$	$2.8 \cdot 10^{-3} \pm 0 \cdot 10^{-4}$
std_{IC}: Training ICs on $\boldsymbol{x}$	$4.1 \cdot 10^{-4} \pm 4 \cdot 10^{-5}$	$5.5 \cdot 10^{-4} \pm 6 \cdot 10^{-5}$
std_{IC}: Training ICs on $\boldsymbol{v}$	$8.3 \cdot 10^{-4} \pm 8 \cdot 10^{-5}$	$1.20 \cdot 10^{-3} \pm 9 \cdot 10^{-5}$
mean_{IC}: Random ICs on $\boldsymbol{x}$	$1.5 \cdot 10^{-3} \pm 1 \cdot 10^{-4}$	$1.8 \cdot 10^{-3} \pm 1 \cdot 10^{-4}$
mean_{IC}: Random ICs on $\boldsymbol{v}$	$2.6 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$	$2.7 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$
std_{IC}: Random ICs on $\boldsymbol{x}$	$4.1 \cdot 10^{-4} \pm 3 \cdot 10^{-5}$	$5.5 \cdot 10^{-4} \pm 4 \cdot 10^{-5}$
std_{IC}: Random ICs on $\boldsymbol{v}$	$8.4 \cdot 10^{-4} \pm 6 \cdot 10^{-5}$	$1.2 \cdot 10^{-3} \pm 1 \cdot 10^{-4}$

Table 3.14: (CS) Trajectory Errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory errors in $\boldsymbol{x}/\boldsymbol{v}$ is calculated using (3.4.3)/(3.4.4).

We consider the Flocking score (at $t = T_f$) taken from [46],

$$I_{\text{flock}} = \sum_{1=i<i'=N} \|\boldsymbol{v}_i(T_f) - \boldsymbol{v}_{i'}(T_f)\|$$

When $I_{\text{flock}} = 0$, perfect flocking occurs; however we use $I_{\text{flock}} < 0.1$ to indicate flocking.

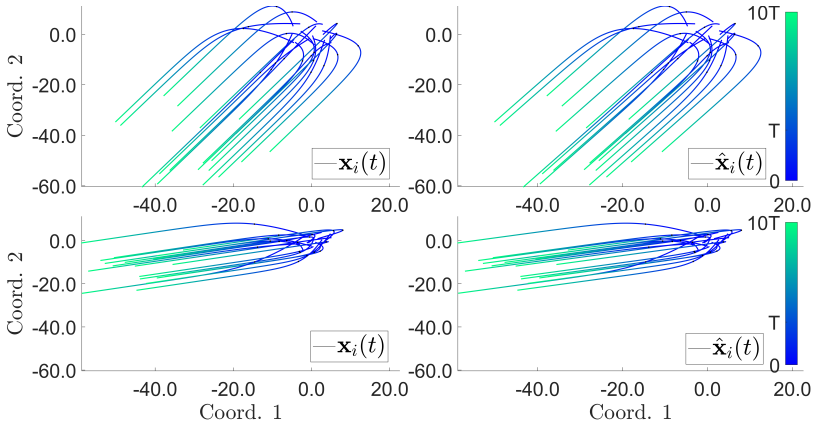


Figure 3.4: (CS) Comparison of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, with the errors reported in table 3.14. The first row of trajectories are generated from an initial condition taken from the observation data. The second row of trajectories are generated from another randomly chosen initial condition. The first column of trajectories are generated from the true interaction kernel, whereas the second column of trajectories are generated from our estimated kernel with the same initial conditions. The color of the trajectory indicates the flow of time, from deep blue (at $t = T_0$) to light green (at $t = T_f$).

The prediction capability of the estimated systems in the form of confusion matrix is reported in table 3.15.

	Predicted Non-Flocking	Predicted Flocking
True Non-Flocking: Training ICs	$0.0 \pm 0.1\%$	0%
True Flocking: Training ICs	$0.0 \pm 0.1\%$	$99.9 \pm 0.2\%$
True Non-Flocking: Random ICs	$0.1 \pm 0.2\%$	0%
True Flocking: Random ICs	0%	$99.9 \pm 0.2\%$

Table 3.15: (CS) Confusion Matrix: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). It is generated using table 3.4.

With $\beta = \frac{1}{4} < \frac{1}{2}$, the true system is guaranteed to show flocking. Since we have no control over when the flocking would occur, we use I_{flock} to help us capture the essence of flocking behavior, i.e. $I_{\text{flock}} = 0$ (or close to 0). And our estimated system can capture the same behavior (flocking or not) with high probability. We provide more statistics about the confusion matrix in order to understand our prediction of clustering better in table 3.16.

	Accuracy	Precision	Recall	F-Score
Training ICs	$100.0 \pm 0.1\%$	100%	$100.0 \pm 0.1\%$	$99.98 \pm 0.06\%$
Random ICs	100%	100%	100%	100%

Table 3.16: (CS) Confusion Matrix Statistics: ICs used in the training set, new ICs randomly drawn from $\mu^{\boldsymbol{x}}$. It is generated using table 3.5.

In order for us to provide more quantitative insight into the predication capability of our estimator for the case of flocking, we consider two different pattern indicator scores. First, PI_1 is the relative error of I_{flock} between true and predicted systems, averaged over M trials. Second, we consider another important quantity, the center of mass velocity, $\boldsymbol{v}_{\text{CM}}$. It is given by $\boldsymbol{v}_{\text{CM}} = \frac{\sum_{i=1}^{N} m_i \boldsymbol{v}_i(T_f)}{\sum_{i=1}^{N} m_i}$. In the case of the CS dynamics ($m_i = 1$ for $i = 1, \cdots, N$), $\boldsymbol{v}_{\text{CM}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}_i(T_f)$. Then we define PI_2 to be the relative error of the predicted center of mass velocity and true center of mass velocity averaged over M trials. The scores are reported in table 3.17.

	PI_1	PI_2
mean_{IC}: Training ICs	$1.13 \cdot 10^{-2} \pm 8 \cdot 10^{-4}$	$6 \cdot 10^{-15} \pm 3 \cdot 10^{-15}$
std_{IC}: Training ICs	$5.1 \cdot 10^{-3} \pm 7 \cdot 10^{-4}$	$3 \cdot 10^{-14} \pm 3 \cdot 10^{-14}$
mean_{IC}: Random ICs	$1.12 \cdot 10^{-2} \pm 7 \cdot 10^{-4}$	$7 \cdot 10^{-15} \pm 4 \cdot 10^{-15}$
std_{IC}: Random ICs	$5.4 \cdot 10^{-3} \pm 7 \cdot 10^{-4}$	$4 \cdot 10^{-14} \pm 6 \cdot 10^{-14}$

Table 3.17: (CS) Pattern Indicator Scores: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows).

As it is shown in table 3.17, our estimated system can predict I_{flock} with relatively high accuracy. Surprisingly, our estimated system can reproduce $\boldsymbol{v}_{\text{CM}}$ down to numerical accuracy.

3.5.3 Fish Milling in 2 dimensions

Next we consider a more complicated cohesive collective system: a dynamical system which produces milling patterns, where each agent rotates around the same axis or about the same point. The models we consider have been proposed in [38, 37] (see references therein for a variety of sources for the biological roots of these models).

Useful background references for the two-dimensional models are [92, 1] as well as the primer [11]. Further theoretical study of models of this type has been done in [29, 54, 5].

The governing equations of the Fish Milling Dynamics in $\mathbb{R}^2$ (FM2D) of [38] are,

$$m_i \ddot{\boldsymbol{x}}_i = (\alpha - \beta \left\| \dot{\boldsymbol{x}}_i \right\|^2) \dot{\boldsymbol{x}}_i - \nabla_{\boldsymbol{x}_i} U_i, \quad \text{for } i = 1, \cdots N. \tag{3.5.1}$$

Here, U_i is the Morse potential describing the interaction of the i^{th} agent with the other agents in the system, defined as follows

$$U_i = \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \big(- C_a \mathrm{e}^{-\frac{\left\| \boldsymbol{x}_i - \boldsymbol{x}_{i'} \right\|}{\ell_a}} + C_r \mathrm{e}^{-\frac{\left\| \boldsymbol{x}_i - \boldsymbol{x}_{i'} \right\|}{\ell_r}} \big).$$

Here C_a/C_r are the attraction/repulsion strengths and ℓ_a/ℓ_r are the effective attraction/repulsion lengths. Table 3.18 shows how the FM2D dynamics fits into the framework of (3.2.2).

Category	ξ_i	K	$\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i)$	ϕ^A	$s^{\boldsymbol{x}}_{i,i'}$	$\phi^E(r)$
Value	$\emptyset$	1	$(\alpha - \beta \left\| \dot{\boldsymbol{x}}_i \right\|^2) \dot{\boldsymbol{x}}_i$	$\emptyset$	$\emptyset$	$\frac{N}{r}\big(\frac{C_a}{\ell_a}\mathrm{e}^{-\frac{r}{\ell_a}} - \frac{C_r}{\ell_r}\mathrm{e}^{-\frac{r}{\ell_r}}\big)$

Table 3.18: (FM2D) Mapping to (3.2.2)

M	d	α	β	T_f	T	$\mu^{\boldsymbol{x}}$	$\mu^{\dot{\boldsymbol{x}}}$
500	2	1.6	0.5	20	4	Unif. on $[0,1]^2$	Unif. on $[0,0]^2$

Table 3.19: (FM2D) Parameters for Experiment Setup

The delicate balance between the self-propelling force produced by $\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i)$ and the collective force induced by the energy kernel U_i can create a wide range of patterns for different initial conditions. Milling patterns (single or double milling) are one of the most interesting ones. Unlike the well-understood Cucker-Smale model, necessary and sufficient conditions on the interaction kernels and ICs that guarantee the existence milling patterns seem to be unknown. These milling patterns result from the balance

of the non-collective force and the collective force induced by the energy kernel U_i (especially when U_i is not H-stable, double-milling would occur, see Fig. 1 in [38]), and are therefore rather sensitive to the selection of parameters: relatively small differences in the interaction laws can correspond to dynamical systems with very different dynamical patterns. The estimator error between the true and estimated interaction kernel may therefore offer little insight information into how well our estimated dynamics can re-produce milling patterns at large time. The confusion matrix and pattern indicator scores are finer indicators of performance in this case.

We consider the following interaction law,

$$\phi^E(r) = \frac{N}{r}\left(\mathrm{e}^{-\frac{r}{2}} - 2\mathrm{e}^{-\frac{r}{0.5}}\right),$$

With this setup, a double-milling pattern appears 100% of the time (see [37]). The other parameters are reported in table 3.19. We use piecewise constant polynomials with $n^{\boldsymbol{x}} = 122$ basis functions to approximate ϕ^E. The comparison of the true ϕ^A and the estimated $\hat{\phi}^A$ is shown in Fig.3.5. As it is shown in Fig. 3.5, our estimator closely

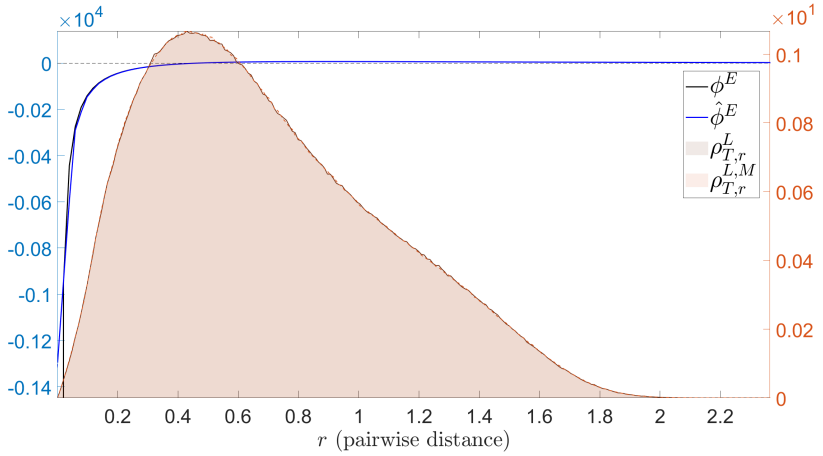


Figure 3.5: (FM2D) Comparison of ϕ^E and $\hat{\phi}^E$, with the relative error being $6.0 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$ (calculated using (3.4.2)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in blue dotted lines. Shown in the background is the comparison of approximated ρ^L_T versus the empirical $\rho^{L,M}_T$.

resembles ϕ^E, however when r is close to 0, there is a sharp drop of ϕ^E to $-\infty$, the availability of r data close to 0 becomes scarcer, and since we are using a uniform basis to approximate ϕ^E, the difference between ϕ^E and $\hat{\phi}^E$ is apparent in this range. The comparison of the true trajectory $\boldsymbol{X}(t)$ and learned $\hat{\boldsymbol{X}}(t)$ is shown in Fig. 3.6. Our predicted system can still estimate the position/velocity of the agents in large

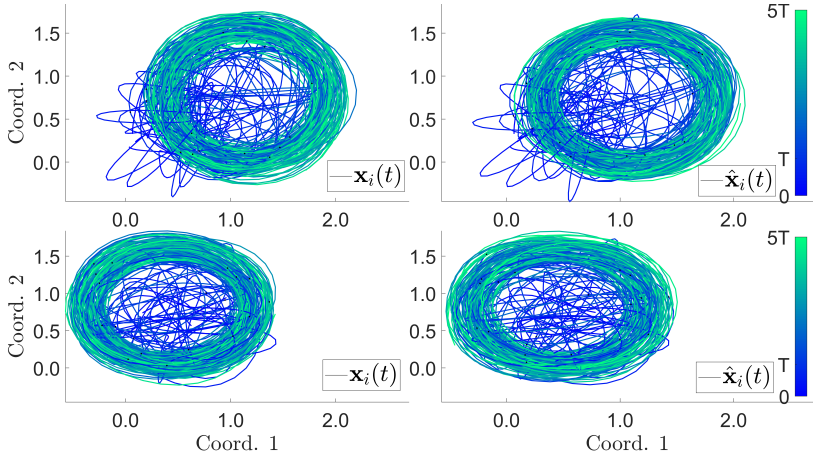


Figure 3.6: (FM2D) Comparison of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, with the errors reported in table 3.20. The first row of trajectories are generated from an initial condition taken from the observation data. The second row of trajectories are generated from another randomly chosen initial condition. The first column of trajectories are generated from the true interaction kernel, whereas the second column of trajectories are generated from our estimated kernel with the same initial conditions. The color of the trajectory indicates the flow of time, from deep blue (at $t = T_0$) to light green (at $t = T_f$).

time, i.e., for $t \gg T$, with relatively small error, around 10^{-2}. Moreover, when the dynamics enters its milling state, our predicted system is also in the same milling state. We provide a quantitative insight into the difference between trajectories in table 3.20.

	$[0,T]$	$[T,T_f]$
mean_{IC}: Training ICs on $\boldsymbol{x}$	$2.35\cdot10^{-1}\pm8\cdot10^{-3}$	$8.38\cdot10^{-1}\pm9\cdot10^{-3}$
mean_{IC}: Training ICs on $\boldsymbol{v}$	$3.6\cdot10^{-1}\pm1\cdot10^{-2}$	$1.13\pm1\cdot10^{-2}$
std_{IC}: Training ICs on $\boldsymbol{x}$	$1.18\cdot10^{-1}\pm7\cdot10^{-3}$	$2.59\cdot10^{-1}\pm9\cdot10^{-3}$
std_{IC}: Training ICs on $\boldsymbol{v}$	$1.64\cdot10^{-1}\pm6\cdot10^{-3}$	$3.1\cdot10^{-1}\pm1\cdot10^{-2}$
mean_{IC}: Random ICs on $\boldsymbol{x}$	$2.34\cdot10^{-1}\pm8\cdot10^{-3}$	$8.35\cdot10^{-1}\pm9\cdot10^{-3}$
mean_{IC}: Random ICs on $\boldsymbol{v}$	$3.6\cdot10^{-1}\pm1\cdot10^{-2}$	$1.12\pm1\cdot10^{-2}$
std_{IC}: Random ICs on $\boldsymbol{x}$	$1.15\cdot10^{-1}\pm5\cdot10^{-3}$	$2.5\cdot10^{-1}\pm1\cdot10^{-2}$
std_{IC}: Random ICs on $\boldsymbol{v}$	$1.63\cdot10^{-1}\pm8\cdot10^{-3}$	$3.1\cdot10^{-1}\pm1\cdot10^{-2}$

Table 3.20: (FM2D) Trajectory Errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory errors in $\boldsymbol{x}/\boldsymbol{v}$ is calculated using (3.4.3)/(3.4.4).

We are getting 10^{-1} relative accuracy for estimating positions/velocities within the learning time interval (i.e. $[0,T]$); however, as time goes on, we can see roughly linear growth of the errors ($T_f = 5T$).

We consider the center of mass position $\boldsymbol{x}_{\text{CM}}(t) = \frac{\sum_{i=1}^N m_i \boldsymbol{x}_i(t)}{\sum_{i=1}^N m_i}$. In the case of $m_i = 1$ for FM2D, it becomes, $\boldsymbol{x}_{\text{CM}}(t) = \frac{1}{N}\boldsymbol{x}_i(t)$. We consider the indicator score I_s (at $t = T_f$), $I_s = I_{\text{flock}} - I_{\text{mill}}$, where $I_{\text{flock}}, I_{\text{mill}}$ are taken from [38]. Here, the flocking score I_{flock} is defined as,

$$I_{\text{flock}} = \left\| \frac{\sum_{i=1}^N \boldsymbol{v}_i(T_f)}{\sum_{i=1}^N \|\boldsymbol{v}_i(T_f)\|} \right\|.$$

Again $I_{\text{flock}} = 1$ when perfect flocking occurs. Then we consider the the milling score I_{mill} as follows,

$$I_{\text{mill}} = \left\| \frac{\sum_{i=1}^N \|(\boldsymbol{x}_i(T_f) - \boldsymbol{x}_{\text{CM}}(T_f)) \times \boldsymbol{v}_i(T_f)\|}{\sum_{i=1}^N \|\boldsymbol{x}_i(T_f) - \boldsymbol{x}_{\text{CM}}(T_f)\| \, \|\boldsymbol{v}_i(T_f)\|} \right\|.$$

When $I_{\text{mill}} = 1$ when perfect milling[4] (around the same axis) occurs; meanwhile $I_{\text{flock}} = 0$ if $I_{\text{mill}} = 1$. Therefore $I_s \in [-1, 1]$. As suggested by the thresholds in [37], we use the case, $I_s \leq -0.5$, to indicate milling. The true systems always show milling pattern (in fact, it shows double milling), and 100% of our estimated systems also

[4]The reason why we choose to use $\|(\boldsymbol{x}_i(T_f) - \boldsymbol{x}_{\text{CM}}(T_f)) \times \boldsymbol{v}_i(T_f)\|$ is because it covers the case of double milling: half of the agents are rotating around the same axis, and the other half rotate around the same axis but in the exact opposite direction.

show milling.

Next for the pattern indicator scores, let PI_1 be the relative error of I_s over M trials. And PI_2 is the relative error between the pair $(I_{\text{flock}}, I_{\text{mill}})$ (in ℓ_2 norm) over M trials. The scores are reported in table 3.21. Milling patterns in dynamics are very

	PI_1	PI_2
$\mathrm{mean}_{\mathrm{IC}}$: Training ICs	$2.47 \cdot 10^{-2} \pm 6 \cdot 10^{-4}$	$2.21 \cdot 10^{-2} \pm 5 \cdot 10^{-4}$
$\mathrm{std}_{\mathrm{IC}}$: Training ICs	$2.2 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$	$1.8 \cdot 10^{-2} \pm 1 \cdot 10^{-3}$
$\mathrm{mean}_{\mathrm{IC}}$: Random ICs	$2.49 \cdot 10^{-2} \pm 3 \cdot 10^{-4}$	$2.22 \cdot 10^{-2} \pm 3 \cdot 10^{-4}$
$\mathrm{std}_{\mathrm{IC}}$: Random ICs	$2.26 \cdot 10^{-2} \pm 6 \cdot 10^{-4}$	$1.83 \cdot 10^{-2} \pm 6 \cdot 10^{-4}$

Table 3.21: (FM2D) Pattern Indicator Scores: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows).

delicate. The intricate balance α/β and the H-stability of U_i decides the appearance of milling in a dynamics, especially when U_i is not H-stable for double milling patterns. In the case of the true dynamics showing milling (to be exact, double milling), our predicted systems can capture the same behavior (with high accuracy) both in terms of I_s and the pair, $(I_{\text{flock}}, I_{\text{mill}})$.

3.5.4 Fish Milling in 3 dimensions

Next, we consider a cohesive collective dynamics in $3D$ of self-propelled particles within a fluid environment, introduced in [37]. It is a more complicated $3D$ extension of the FM2D model, where agents could experience self-propelling force in a fluid.

The governing equations of the Fish Milling Dynamics in $\mathbb{R}^3$ (FM3D) are,

$$\ddot{\boldsymbol{x}}_i = -\gamma(\dot{\boldsymbol{x}}_i - \boldsymbol{u}(\boldsymbol{x}_i)) + \mathbf{F}_M(\dot{\boldsymbol{x}}_i, \boldsymbol{u}(\boldsymbol{x}_i)) - \nabla_{\boldsymbol{x}_i} U_i, \quad \text{for } i = 1, \cdots N.$$

Here, $\boldsymbol{u}$ is the lab-frame fluid velocity generated at position $\boldsymbol{x}_i$, $-\gamma(\dot{\boldsymbol{x}}_i - \boldsymbol{u}(\boldsymbol{x}_i))$ gives the drag force ($\gamma > 0$), $\mathbf{F}_M(\dot{\boldsymbol{x}}_i, \boldsymbol{u}(\boldsymbol{x}_i))$ represents the self-propelling motility force, and $-\nabla_{\boldsymbol{x}_i} U_i$ is the agent-to-agent interaction force on agent i, and the energy potential U_i

is the same Morse potential defined in sec. 3.5.3. $\mathbf{F}_M$ is defined as follows

$$\mathbf{F}_M(\dot{\boldsymbol{x}}_i, \boldsymbol{u}(\boldsymbol{x}_i)) = (\alpha - \beta \left\|\dot{\boldsymbol{x}}_i - \lambda \boldsymbol{u}(\boldsymbol{x}_i)\right\|^2)(\dot{\boldsymbol{x}}_i - \lambda \boldsymbol{u}(\boldsymbol{x}_i)).$$

The parameters, $\alpha, \beta > 0$, give the self-acceleration and deceleration, respectively; $0 \leq \lambda \leq 1$ is a perception coefficient, with $\lambda = 0$ showing a "clear" fluid (and it gives the classical Rayleigh-Helmholtz friction), and $\lambda = 1$ for an "opaque" fluid; and the lab-frame fluid velocity $\boldsymbol{u}$ is given as follows

Table 3.22 shows how the FM3D dynamics fits into the framework of (3.2.2).

Category	ξ_i	K	$\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i)$	ϕ^A	$s^{\boldsymbol{x}}_{i,i'}$	$\phi^E(r)$
Value	$\emptyset$	1	$-\gamma(\dot{\boldsymbol{x}}_i - \boldsymbol{u}(\boldsymbol{x}_i)) + \mathbf{F}_M(\dot{\boldsymbol{x}}_i, \boldsymbol{u}(\boldsymbol{x}_i))$	$\emptyset$	$\emptyset$	$N \cdot \big(\frac{C_a}{r\ell_a}\mathrm{e}^{-\frac{r}{\ell_a}} - \frac{C_r}{r\ell_r}\mathrm{e}^{-\frac{r}{\ell_r}}\big)$

Table 3.22: (FM3D) Mapping to (3.2.2)

M	d	α	β	G	λ	γ	T_f	T	$\mu^{\boldsymbol{x}}$	$\mu^{\dot{\boldsymbol{x}}}$
500	3	10^{-4}	$\frac{\alpha}{3}$	10^{-4}	1.0	10^{-4}	20	4	Unif. on $[0, 2.8\sqrt[3]{3}]^3$	Unif. on $[0, 0]^3$

Table 3.23: (FM3D) Parameters for Experiment Setup

The delicate balance between the self-propelling force (in the presence of a fluid environment) and the collective force induced by the energy kernel U_i can create a wide range of patterns for such dynamics. And the H-stability of U_i is the key at producing milling patterns. Again, we want to understand the milling pattern and predict such a pattern with our estimators when the true system shows milling.

We consider the following interaction law,

$$\phi^E(r) = N \cdot \big(\frac{1.4}{2.8r}\mathrm{e}^{-\frac{2.8r}{1.4}} - \frac{2}{r}\mathrm{e}^{-r}\big).$$

We also use the parameters in table 3.23 to set up the experiment. Piece-wise linear polynomials with $n^{\boldsymbol{x}} = 74$ basis functions are used to approximate ϕ^E. The comparison of the true ϕ^A and the estimated $\hat{\phi}^A$ is shown in Fig.3.7. As it is shown in Fig. 3.7, our

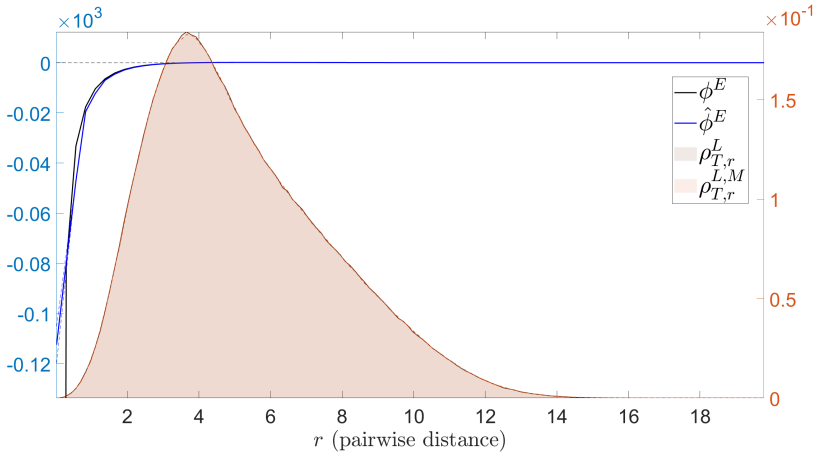


Figure 3.7: (FM3D) Comparison of ϕ^E and $\hat{\phi}^E$, with the relative error being $1.49 \cdot 10^{-1} \pm 3.4 \cdot 10^{-3}$ (calculated using (3.4.2)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in blue dotted lines. Shown in the background is the comparison of approximated ρ^L_T versus the empirical $\rho^{L,M}_T$.

estimator, $\hat{\phi}^E$, deviates from ϕ^E for r close to 0, for similar reasons as those discussed for the $2D$ case. The comparison of the true trajectory $\boldsymbol{X}(t)$ and learned $\hat{\boldsymbol{X}}(t)$ is shown in Fig. 3.8. A $3D$ milling pattern is more complicated than its $2D$ counterpart. In the case of our experiments, some of the trajectories show a pattern of rotation about a fixed point. And our estimated dynamics also shows similar behavior. We provide a quantitative insight into the difference between trajectories in table 3.24.

	$[0,T]$	$[T,T_f]$
mean_{IC}: Training ICs on $\boldsymbol{x}$	$4.9 \cdot 10^{-2} \pm 4 \cdot 10^{-3}$	$6.8 \cdot 10^{-1} \pm 4 \cdot 10^{-2}$
mean_{IC}: Training ICs on $\boldsymbol{v}$	$1.6 \cdot 10^{-1} \pm 1 \cdot 10^{-2}$	$1.2 \pm 3 \cdot 10^{-1}$
std_{IC}: Training ICs on $\boldsymbol{x}$	$5.0 \cdot 10^{-3} \pm 3 \cdot 10^{-4}$	$2.3 \cdot 10^{-1} \pm 3 \cdot 10^{-2}$
std_{IC}: Training ICs on $\boldsymbol{v}$	$3.3 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$	$4 \cdot 10^{-1} \pm 4 \cdot 10^{-1}$
mean_{IC}: Random ICs on $\boldsymbol{x}$	$4.9 \cdot 10^{-2} \pm 4 \cdot 10^{-3}$	$6.8 \cdot 10^{-1} \pm 4 \cdot 10^{-2}$
mean_{IC}: Random ICs on $\boldsymbol{v}$	$1.5 \cdot 10^{-1} \pm 1 \cdot 10^{-2}$	$1.2 \pm 3 \cdot 10^{-1}$
std_{IC}: Random ICs on $\boldsymbol{x}$	$5.1 \cdot 10^{-3} \pm 3 \cdot 10^{-4}$	$2.3 \cdot 10^{-1} \pm 4 \cdot 10^{-2}$
std_{IC}: Random ICs on $\boldsymbol{v}$	$3.2 \cdot 10^{-2} \pm 1 \cdot 10^{-3}$	$4 \cdot 10^{-1} \pm 5 \cdot 10^{-1}$

Table 3.24: (FM3D) Trajectory Errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory errors in $\boldsymbol{x}/\boldsymbol{v}$ is calculated using (3.4.3)/(3.4.4).

We consider the center of mass velocity $\boldsymbol{v}_{\text{CM}}(t) = \frac{\sum_{i=1}^N m_i \boldsymbol{v}_i(t)}{\sum_{i=1}^N m_i}$. In the case of $m_i = 1$

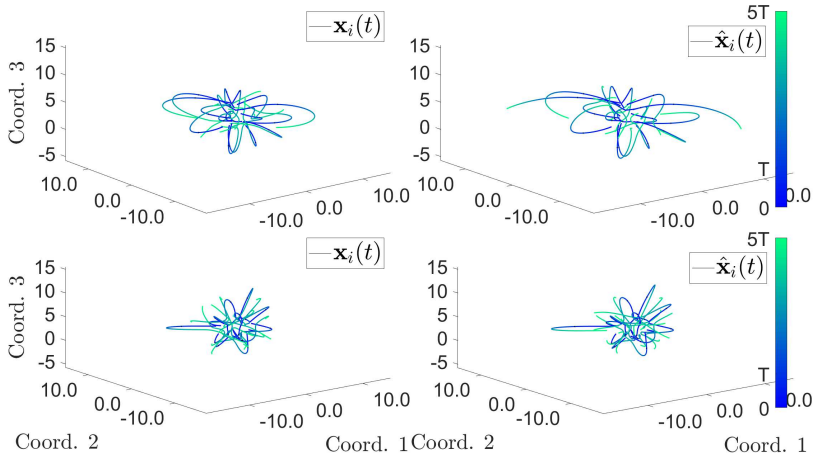


Figure 3.8: (FM3D) Comparison of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, with the errors reported in table 3.24. The first row of trajectories are generated from an initial condition taken from the observation data. The second row of trajectories are generated from another randomly chosen initial condition. The first column of trajectories are generated from the true interaction kernel, whereas the second column of trajectories are generated from our estimated kernel with the same initial conditions. The color of the trajectory indicates the flow of time, from deep blue (at $t = T_0$) to light green (at $t = T_f$).

for FM3D, it becomes, $\boldsymbol{v}_{\text{CM}}(t) = \frac{1}{N}\boldsymbol{v}_i(t)$. We consider the indicator score I_s (at $t = T_f$) from [37],

$$I_s = I_{\text{flock}} - I_{\text{mill}}.$$

Here, the flocking score I_{flock} is defined as,

$$I_{\text{flock}} = 1 - \frac{\sum_{i=1}^{N} \|\boldsymbol{v}_i(T_f) - \boldsymbol{v}_{\text{CM}}(T_f)\|}{N\sqrt{\frac{\alpha}{\beta}}}.$$

Again $I_{\text{flock}} = 1$ when perfect flocking occurs. The milling score I_{mill} has two pieces: we first define the rotational axis ω_i for agent i,

$$\omega_i = \frac{\boldsymbol{v}_i(T_f) \times (m_i \dot{\boldsymbol{v}}_i(T_f))}{\|\boldsymbol{v}_i(T_f)\| \, \|m_i \dot{\boldsymbol{v}}_i(T_f)\|};$$

next,

$$I_{\text{mill}} = \frac{\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \langle \omega_i, \omega_j \rangle}{N(N-1)}.$$

When $I_{\text{mill}} = 1$ when perfect milling (around the same axis) occurs; meanwhile $I_{\text{flock}} = 0$ if $I_{\text{mill}} = 1$. Therefore $I_s \in [-1, 1]$. As suggested by the thresholds in [37], we use the case, $I_s \leq -0.5$, to indicate milling, see the results in table 3.25.

	Predicted Non-Milling	Predicted Milling
True Non-Milling: Training ICs	$99 \pm 2\%$	$1 \pm 2\%$
True Milling: Training ICs	0%	0%
True Non-Milling: Random ICs	$99 \pm 3\%$	$1 \pm 3\%$
True Milling: Random ICs	0%	0%

Table 3.25: (FM3D) Confusion Matrix: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). It is generated using table 3.4.

The true FM3D systems show a non-milling[5] pattern. Furthermore, our predicted systems show a exceptionally similar probability of displaying the non-milling patterns.

We provide more statistics about the confusion matrix in order to understand our prediction of milling behavior better in table 3.26.

	Accuracy	Precision	Recall	F-Score
Training ICs	$99 \pm 2\%$	$30 \pm 48\%$	100%	$30 \pm 48\%$
Random ICs	$99 \pm 3\%$	$70 \pm 48\%$	100%	$70 \pm 48\%$

Table 3.26: (FM3D) Confusion Matrix Statistics: ICs used in the training set, new ICs randomly drawn from $\mu^{\boldsymbol{x}}$. It is generated using table 3.5.

Next, we use the pattern indicator scores to probe deeper into the actual large-time behavior of our predicted systems. PI_1 is the relative error of I_s. And PI_2 is the relative error of predicting the pair $(I_{\text{flock}}, I_{\text{mill}})$ (in ℓ_2 norm). The scores are reported in table 3.27.

	PI_1	PI_2
mean_{IC}: Training ICs	$3 \cdot 10^{-1} \pm 1 \cdot 10^{-1}$	$3 \cdot 10^{-1} \pm 1 \cdot 10^{-1}$
std_{IC}: Training ICs	$3 \cdot 10^{-1} \pm 2 \cdot 10^{-1}$	$3 \cdot 10^{-1} \pm 3 \cdot 10^{-1}$
mean_{IC}: Random ICs	$3 \cdot 10^{-1} \pm 1 \cdot 10^{-1}$	$3 \cdot 10^{-1} \pm 1 \cdot 10^{-1}$
std_{IC}: Random ICs	$3 \cdot 10^{-1} \pm 3 \cdot 10^{-1}$	$3 \cdot 10^{-1} \pm 3 \cdot 10^{-1}$

Table 3.27: (FM3D) Pattern Indicator Scores: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows).

[5]The milling score given by [37] fails to capture the case of milling around a fixed point, hence the non-milling pattern.

Not only can we predict I_s with relatively high accuracy, but also can we offer insight into the actual pair of scores, $(I_{\text{flock}}, I_{\text{mill}})$.

3.6 Emergent Behaviors Induced by $\phi(r, s)$

The flexibility of the learning algorithm given in [89] allows for a generalization of the dynamical system where the interaction kernels can depend on more than just one variable, i.e., more than just r (the pairwise distance data). For example, in modeling the movement of groups of animals, field of vision can affect how individuals influence each other; in synchronized fireflies, not only can the fireflies form spatial patterns, their light-emitting states can be also be locked in synchronization. We consider an important example in [102] which models how oscillators can sync and swarm together, hence the interaction kernels depend on both r and ξ (the pairwise difference in phases). Further study of this type of model has been done in [63, 82, 77, 101], a review with applications to computation is given in [100], and a historical review of the development of the synchronization models can be found in [127].

These authors sought to develop a plausible model that could explain systems where a phase or real-valued feature affects the motion – and vice versa. They called such systems "swarmalators" to emphasize the combined behavior of swarming and synchronized oscillation of phases in the system.

For the Synchronized Oscillator Dynamics (SOD), each agent is indexed by i, ξ_i is its phase, $\boldsymbol{x}_i$ is (as usual) its position, ω_i is the fixed natural frequency, $\boldsymbol{v}_i$ is the fixed self-propulsion velocity. The dynamics of $\boldsymbol{x}_i$ and ξ_i are governed by the following equations,

$$\begin{cases} \dot{\boldsymbol{x}}_i &= \boldsymbol{v}_i + \frac{1}{N}\sum_{i'=1}^{N}\left(\frac{\boldsymbol{x}_{i'}-\boldsymbol{x}_i}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|}(A + J\cos(\xi_{i'}-\xi_i)) - B\frac{\boldsymbol{x}_{i'}-\boldsymbol{x}_i}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|^2}\right) \\ \dot{\xi}_i &= \omega_i + \frac{K}{N}\sum_{i'=1}^{N}\frac{\sin(\xi_{i'}-\xi_i)}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|} \end{cases} \tag{3.6.1}$$

Table 3.28 shows how the SOD dynamics fits into the framework of (3.2.1).

Category	K	$\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \xi_i)$	$\mathbf{F}^{\xi}(\boldsymbol{x}_i, \xi_i)$	$\phi^A(r, s^{\boldsymbol{x}})$	$\phi^{\xi}(r, s^{\xi})$
Value	1	$\boldsymbol{v}_i$	ω_i	$\frac{A+J\cos(s^{\boldsymbol{x}})}{r} - \frac{B}{r^2}$	$\frac{K\sin(s^{\xi})}{r}$

Table 3.28: Mapping of SOD to (3.2.1)

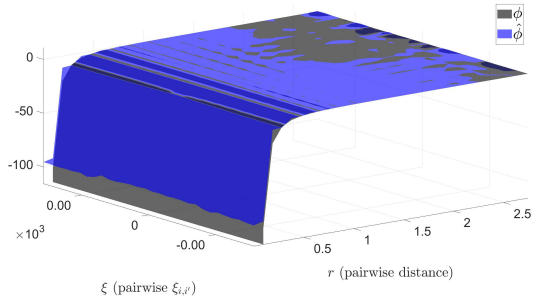
M	$\boldsymbol{v}_i$	ω_i	d	T_f	T	$\mu^{\boldsymbol{x}}$	μ^{ξ}
1000	0	0	4	20	4	Unif. on $[-1,1]^2$	Unif. on $[-\pi,\pi]$

Table 3.29: (SOD) Parameters for Experiment Setup

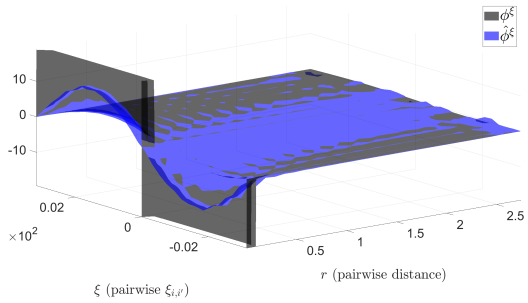
With certain choices of A, J, B and K, the SOD dynamics is going to produce either a static or a non-static spatial pattern with either phases in sync or out of sync (a total of 5 different states, see [102] for details). We consider the following interaction law,

$$\phi^E(r, s^{\boldsymbol{x}}) = \frac{1 + J\cos(s^{\boldsymbol{x}})}{r} - \frac{1}{r^2} \quad \text{and} \quad \phi^{\xi} = \frac{K\sin(s^{\xi})}{r}.$$

Here K and J are changing and we take $s^{\boldsymbol{x}} = s^{\xi} = \xi$ (the pairwise difference in the phases, i.e., $\xi_{i'} - \xi_i$). We consider a particular set of (J, K) values, i.e. $(J, K) = (0.1, 1)$, which gives a static synchronous state. In table 3.29 we describe the other parameters that we use to set up the experiment. Here we use piecewise linear polynomials with $n^{\boldsymbol{x}} = 900$ (with 30 basis functions in each dimension) and $n^{\xi} = 900$ (with 30 basis functions in each dimension) basis functions to approximate ϕ^E and ϕ^{ξ}. The comparison of the true ϕ^A and the estimated $\hat{\phi}^A$ is shown in Fig.3.9. As is shown in Fig. 3.9(a) and Fig. 3.9(b), even with the interaction laws being 2-dimensional, we can still infer from the data with around 10^{-1} relative accuracy with relatively small number of basis functions. A comparison of the true trajectory $\boldsymbol{X}(t)$ and learned $\hat{\boldsymbol{X}}(t)$ is shown in Fig. 3.10. A visual comparison of the trajectories between the true and estimated dynamics shows that the difference is small. An more quantitative insight into the difference between trajectories is provided in table 3.30.



((a)) Comparison of ϕ^E and $\hat{\phi}^E$, the relative error is $4.5 \cdot 10^{-1} \pm 9 \cdot 10^{-2}$ (calculated using (3.4.2)).



((b)) Comparison of ϕ^ξ and $\hat{\phi}^\xi$, the relative error is $2 \cdot 10^{-1} \pm 1 \cdot 10^{-1}$ (calculated using (3.9.2)).

Figure 3.9: (SOD) The true interaction laws are shown in black, and the mean estimated interaction laws are shown in blue.

	$[0, T]$	$[T, T_f]$
mean_{IC}: Training ICs on $\boldsymbol{x}$	$4.69 \cdot 10^{-2} \pm 5 \cdot 10^{-4}$	$7.2 \cdot 10^{-2} \pm 1 \cdot 10^{-3}$
mean_{IC}: Training ICs on ξ	$2.97 \cdot 10^{-2} \pm 5 \cdot 10^{-4}$	$3 \cdot 10^{-1} \pm 7 \cdot 10^{-1}$
std_{IC}: Training ICs on $\boldsymbol{x}$	$9 \cdot 10^{-3} \pm 1 \cdot 10^{-3}$	$3.5 \cdot 10^{-2} \pm 1 \cdot 10^{-3}$
std_{IC}: Training ICs on ξ	$3.0 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$	8 ± 22
mean_{IC}: Random ICs on $\boldsymbol{x}$	$4.67 \cdot 10^{-2} \pm 6 \cdot 10^{-4}$	$7.2 \cdot 10^{-2} \pm 1 \cdot 10^{-3}$
mean_{IC}: Random ICs on ξ	$3.0 \cdot 10^{-2} \pm 1 \cdot 10^{-3}$	$1.2 \cdot 10^{-1} \pm 8 \cdot 10^{-2}$
std_{IC}: Random ICs on $\boldsymbol{x}$	$9 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$3.5 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
std_{IC}: Random ICs on ξ	$2.9 \cdot 10^{-2} \pm 3 \cdot 10^{-3}$	2 ± 3

Table 3.30: (SOD) Trajectory Errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory errors in $\boldsymbol{x}/\xi$ is calculated using (3.4.3)/(3.4.5).

The Synchronized Oscillator dynamics is a complex dynamical system with $\boldsymbol{x}_i$ and a periodic ξ_i interacting with each other within the agents themselves and also collectively among the other agents; however we are still able to maintain 2-digit relative accuracy in reproducing the trajectories, and in predicting the large-time behavior of the trajectories, the errors do not grow exponentially. We are considering the static synchronous state, hence we check the distribution of the phases at time T_f to see if the phases are in sync. In particular, we use the mean and the variance of the phases at time T_f. If the variance of $\{\xi_i(T_f)\}_{i=1}^N$ is smaller than 0.01, we say that the dynamics is in static synchronous state. Since the true systems always show synchronous state, our estimated systems also shows synchronization of phases with

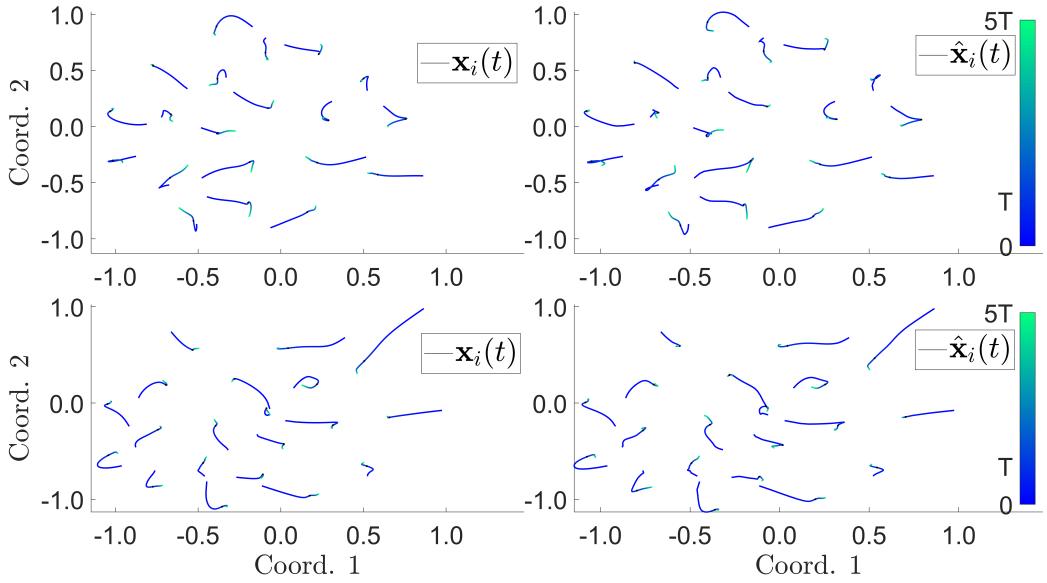


Figure 3.10: (SOD) Comparison of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, with the errors reported in table 3.30. The first row of trajectories are generated from an initial condition taken from the observation data. The second row of trajectories are generated from another randomly chosen initial condition. The first column of trajectories are generated from the true interaction kernel, whereas the second column of trajectories are generated from our estimated kernel with the same initial conditions. The color of the trajectory indicates the flow of time, from deep blue (at $t = T_0$) to light green (at $t = T_f$).

the same certainty.

Next, we use the following pattern indicator scores to discuss the quantitative predication performance of our estimators. PI_1 is the relative error of the variance of the phases at time T_f, and PI_2 is the relative error of the mean of the phases at time T_f. The scores are reported in table 3.31.

	PI_1	PI_2
$\mathrm{mean}_{\mathrm{IC}}$: Training ICs	0	$8 \cdot 10^{-2} \pm 2 \cdot 10^{-1}$
$\mathrm{std}_{\mathrm{IC}}$: Training ICs	0	2 ± 5
$\mathrm{mean}_{\mathrm{IC}}$: Random ICs	0	$3 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$
$\mathrm{std}_{\mathrm{IC}}$: Random ICs	0	$4 \cdot 10^{-1} \pm 6 \cdot 10^{-1}$

Table 3.31: (SOD) Pattern Indicator Scores: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows).

Our estimated systems display exactly the same synchronization behavior as the true system (variance of the phases is exactly 0). Meanwhile, we can reproduce the

final synchronized phase with relative high accuracy, however it comes with a big uncertainty.

3.7 Emergent Behaviors Induced by Parametric Families of Interaction Kernels

In recent years, there has been rapid growth in developing algorithms (either theoretical or numerical) to identify the governing equations of physical systems based on observed data. A notable collection of these approaches assume a parametric form of the equations to perform various kinds of regression, usually sparse regression against an enormous library of standard mathematical functions, to fit the parameters [23, 115, 22]. Other approaches use force-based, statistical mechanics, and multiscale methods – see the works [7, 91, 74, 14, 118]. Our non-parametric learning approach can also be used to discover the elaborate structure of the true interaction law, i.e., $\phi(r) = \phi(r; \boldsymbol{P})$, where $\boldsymbol{P} = \begin{bmatrix} p_1 & \cdots & p_k \end{bmatrix}$ is a vector of parameters. In many settings, ϕ can be written as $\phi(r; \boldsymbol{P}) = J(\boldsymbol{P})\phi_m(r)$, where $J(\cdot)$ might offer physical insight through its effect on the parameters. In this chapter, we will focus on the case when $\boldsymbol{P}$ is 1-dimensional, i.e., a family of one-parameter interaction laws.

We consider a simplified planetary movement in our solar system (GSS) as a second order collective dynamical system example with parametric interaction laws. We take $\boldsymbol{x}_i(t) \in \mathbb{R}^2$ or $\mathbb{R}^3$ as the position of each planet (only the planets in the inner-solar system are considered, i.e., Mercury, Venus, Earth and Mars, hence $N = 5$). Their positions are governed by the following form of Newton's Law,

$$\tilde{m}_i^{\mathcal{I}} \ddot{\boldsymbol{x}}_i(t) = \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \frac{G \tilde{m}_i^{\mathcal{G}} \tilde{m}_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|^2} \cdot \frac{\boldsymbol{x}_{i'} - \boldsymbol{x}_i}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|}, \quad \text{for } i = 1, \cdots, N. \tag{3.7.1}$$

$\tilde{m}_i^{\mathcal{I}}$ is the inertia mass of the i^{th} astronomical object (AO), and $\tilde{m}_i^{\mathcal{G}}$ is the gravitational mass of the corresponding AO. In our setting we will assume that they are the same, hence (3.7.1) can be simplified to,

$$\ddot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{G\tilde{m}_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|^3}(\boldsymbol{x}_{i'} - \boldsymbol{x}_i), \quad \text{for } i = 1, \cdots, N. \tag{3.7.2}$$

Here $\tilde{m}_i$ is the unknown mass of the i^{th} AO, and $G = 6.67408 \cdot 10^{-11}\text{m}^3\text{kg}^{-1}\text{s}^{-2}$ is the gravitational constant (known to the algorithm). There are a total of 5 different types of agents (each AO is of its own type) in this system, and the true interaction laws are

$$\phi^E_{k,k'}(r; \tilde{m}_{k'}) = G\tilde{m}_{k'} \cdot \frac{1}{r^3}, \quad \text{for } k, k' = 1, \cdots, 5.$$

Here, the $\phi^E_{k,k'}$ is parameterized by $J(p_1) = Gp_1$ with $p_1 = \tilde{m}_{k'}$. Table 3.32 shows how the GSS dynamics fits into the framework of (3.2.2).

Category	m_i	ξ_i	K	$s^{\boldsymbol{x}}_{i,i'}$	$\mathbf{F}^{\boldsymbol{x}}(\boldsymbol{x}_i, \dot{\boldsymbol{x}}_i)$	$\phi^E_{k,k'}(r)$	ϕ^A
Value	1	$\emptyset$	N	$\emptyset$	$\emptyset$	$\frac{G\tilde{m}_{k'}}{r^3}$	$\emptyset$

Table 3.32: (GSS) Mapping (3.2.2)

M	d	T_f	T
500	2	913day	182.6day

Table 3.33: (GSS) Parameters for Experiment Setup

We also use the parameters in table 3.33 to set up the experiment. These parameters are based on simple astronomical features of the system and are used for simulation of the dynamics and getting an appropriate and realistic number of observations. We use piecewise linear polynomials with $n^E_{k,k'} = 100$ for $\phi^E_{k,k'}$ when $k \neq k'$, and piecewise constant polynomials with $n^E_{k,k} = 1$ for $\phi^E_{k,k}$. Each AO is given an index as follows: the Sun is assigned an index 1, and depending on the distance from the Sun, the index is increased gradually, and stopping at 5 for Mars. We use the following units: 1 day for time scale, 10^6 km for length scale, and 10^{24} kg for mass scale. The gravitational

constant G becomes $8.64^2 \cdot 6.67408 \cdot 10^{-6}(10^6\text{km})^3(10^{24}\text{kg})^{-1}\text{day}^{-2}$. We also use the following data from NASA in table 3.34.

Category	Sun	Mercury	Venus	Earth	Mars
Mass (10^{24}kg)	$1.989 \cdot 10^6$	0.33	4.87	5.97	0.642
Perihelion (10^6km)	N/A	46	107.5	147.1	206.6
Aphelion (10^6km)	N/A	69.9	108.9	152.1	249.2
Orbital Period (day)	N/A	88	224.7	365.2	687

Table 3.34: (GSS) NASA Data for Each AO

The initial position distribution for the astronomical objects, $\mu^{\boldsymbol{x}}$, is constructed as follows: the Sun is always placed at the origin, whereas the planets are randomly placed on ellipses with their corresponding perihelion and aphelion data, and the Sun is sitting at one of the foci (Sun is the common focus of all initial elliptical trajectory). We construct a distribution, $\mu^{\dot{\boldsymbol{x}}}$, which gives the initial velocities for the astronomical objects as follows: the Sun always has zero initial velocity, whereas the planets will have their initial velocity depending on their initial position and satisfying the Vis-Viva equation (see [86] for details). The comparison of the true $\phi^E_{k,k'}$'s and the estimated $\hat{\phi}^E_{k,k'}$'s is shown in Fig.3.11. We inferred a total of $N^2 = 25$ different interaction laws all together from the observation data. As shown in Fig. 3.11, the interactions from planets on the Sun and the Sun on planets are estimated with high accuracy, however the estimated inter-planet interactions offer little valuable insight of the original interactions. This is likely driven by the domination of the sun in terms of effect on the dynamics – due to its mass. The effect of the Sun's mass creates a form of ill-posedness of the system which affects the accuracy of our estimation. Realizing the possibility of a parametric form of the interaction laws, we go through a delicate decoupling procedure detailed in 3.7.1, and produce a cleaned up version of $\hat{\phi}^E_{k,k'}$'s, shown in Fig. 3.12. As shown in Fig. 3.12, we are able to de-noise the original estimators and obtain a much cleaner presentation of the interaction laws. Relative

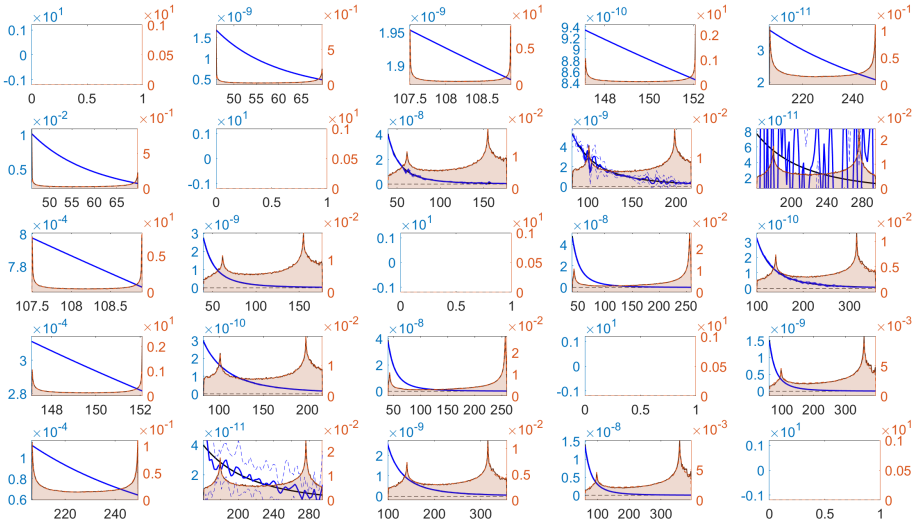


Figure 3.11: (GSS) Comparison of $\phi^E_{k,k'}$'s and $\hat{\phi}^E_{k,k'}$'s , the relative errors are reported in table 3.35. For example, $\phi^E_{1,2}$ on cell $(1,2)$ of the plot represents the true force Mercury has on Sun. Within each sub-plot, the true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in blue dotted lines. Shown in the background of each sub-plot is the comparison of approximated $\rho^{L,k,k'}_{T,r}$ (in lighter color) versus the empirical $\rho^{L,M,k,k'}_{T,r}$ (in darker color).

$L^2(\rho_T)$-errors for each $\phi^E_{k,k'}$ are provided in tables 3.35 and 3.36 in order to re-affirm our claim.

	$k'=1$	$k'=2$	$k'=3$	$k'=4$	$k'=5$
$k=1$	0	$1.6199 \cdot 10^{-4} \pm 7 \cdot 10^{-8}$	$1.13 \cdot 10^{-7} \pm 3 \cdot 10^{-9}$	$7.61 \cdot 10^{-7} \pm 6 \cdot 10^{-9}$	$2.71 \cdot 10^{-5} \pm 2 \cdot 10^{-7}$
$k=2$	$1.6196 \cdot 10^{-4} \pm 8 \cdot 10^{-8}$	0	$1.5 \cdot 10^{-1} \pm 2 \cdot 10^{-2}$	$3.5 \cdot 10^{-1} \pm 7 \cdot 10^{-2}$	9 ± 1.7
$k=3$	$1.03 \cdot 10^{-7} \pm 6 \cdot 10^{-9}$	$3.1 \cdot 10^{-2} \pm 8 \cdot 10^{-3}$	0	$1.42 \cdot 10^{-4} \pm 3 \cdot 10^{-4}$	$10 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$
$k=4$	$7.57 \cdot 10^{-7} \pm 3 \cdot 10^{-9}$	$3 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$	$1.402 \cdot 10^{-2} \pm 7 \cdot 10^{-5}$	0	$2.2 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
$k=5$	$2.717 \cdot 10^{-5} \pm 3 \cdot 10^{-8}$	$8 \cdot 10^{-1} \pm 2 \cdot 10^{-1}$	$3.3 \cdot 10^{-2} \pm 4 \cdot 10^{-3}$	$1.8 \cdot 10^{-2} \pm 3 \cdot 10^{-3}$	0

Table 3.35: (GSS) Relative errors for the estimators, $\hat{\phi}^E_{k,k'}$ (calculated using (3.4.2)).

	$k'=1$	$k'=2$	$k'=3$	$k'=4$	$k'=5$
$k=1$	0	$1.742 \cdot 10^{-4} \pm 4 \cdot 10^{-7}$	$2 \cdot 10^{-5} \pm 2 \cdot 10^{-5}$	$8 \cdot 10^{-5} \pm 8 \cdot 10^{-5}$	$5 \cdot 10^{-3} \pm 4 \cdot 10^{-3}$
$k=2$	$2.9 \cdot 10^{-3} \pm 5 \cdot 10^{-4}$	0	$1.5 \cdot 10^{-1} \pm 2 \cdot 10^{-2}$	$2.4 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$	$5 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$
$k=3$	$9.3 \cdot 10^{-3} \pm 4 \cdot 10^{-4}$	$3.49 \cdot 10^{-2} \pm 2 \cdot 10^{-4}$	0	$3.81 \cdot 10^{-2} \pm 3 \cdot 10^{-4}$	$1.5 \cdot 10^{-1} \pm 9 \cdot 10^{-2}$
$k=4$	$1.7 \cdot 10^{-3} \pm 3 \cdot 10^{-4}$	$2.3 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$	$4.47 \cdot 10^{-2} \pm 4 \cdot 10^{-4}$	0	$1.7 \cdot 10^{-1} \pm 9 \cdot 10^{-2}$
$k=5$	$7 \cdot 10^{-3} \pm 3 \cdot 10^{-3}$	$5 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$	$1.5 \cdot 10^{-1} \pm 9 \cdot 10^{-2}$	$1.7 \cdot 10^{-1} \pm 9 \cdot 10^{-2}$	0

Table 3.36: (GSS) Relative errors for the cleaned-up estimators, $\hat{\phi}^E_{k,k'}$ (calculated using (3.4.2)).

The comparison of true trajectory $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ is shown in Fig. 3.13.

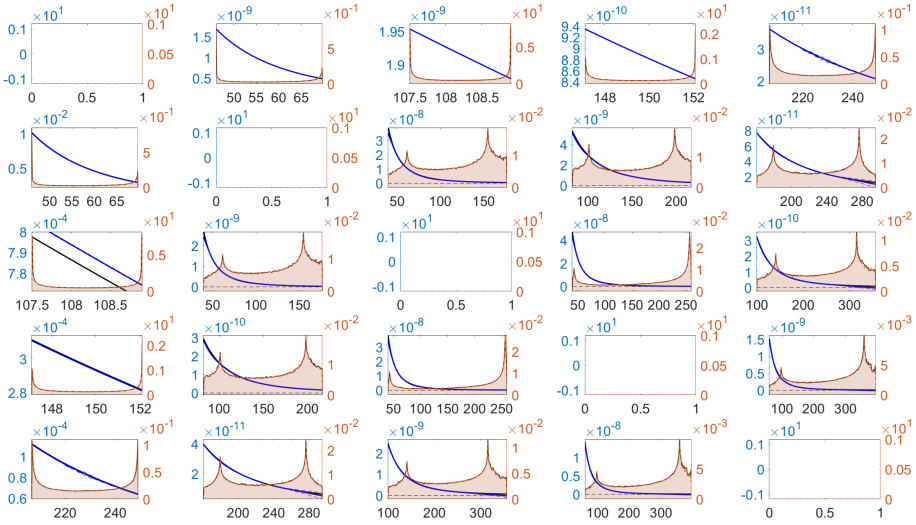


Figure 3.12: (GSS) Comparison of $\phi^E_{k,k'}$'s and cleaned-up $\hat{\phi}^E_{k,k'}$'s , the relative errors are reported in table 3.36. Similar layout and setup as in Fig. 3.11.

	$[0,T]$	$[T,T_f]$
mean_{IC}: Training ICs on $\boldsymbol{x}$	$6.6\cdot10^{-4}\pm2\cdot10^{-5}$	$3.9\cdot10^{-3}\pm2\cdot10^{-4}$
mean_{IC}: Training ICs on $\boldsymbol{v}$	$3.9\cdot10^{-3}\pm1\cdot10^{-4}$	$2.13\cdot10^{-2}\pm8\cdot10^{-4}$
std_{IC}: Training ICs on $\boldsymbol{x}$	$5\cdot10^{-4}\pm1\cdot10^{-4}$	$2.7\cdot10^{-3}\pm4\cdot10^{-4}$
std_{IC}: Training ICs on $\boldsymbol{v}$	$2.5\cdot10^{-3}\pm3\cdot10^{-4}$	$1.30\cdot10^{-2}\pm2\cdot10^{-4}$
mean_{IC}: Random ICs on $\boldsymbol{x}$	$6.8\cdot10^{-4}\pm2\cdot10^{-5}$	$3.9\cdot10^{-3}\pm1\cdot10^{-4}$
mean_{IC}: Random ICs on $\boldsymbol{v}$	$3.9\cdot10^{-3}\pm1\cdot10^{-4}$	$2.13\cdot10^{-2}\pm6\cdot10^{-4}$
mean_{IC}: Random ICs on $\boldsymbol{x}$	$5.3\cdot10^{-4}\pm1\cdot10^{-4}$	$2.5\cdot10^{-3}\pm3\cdot10^{-4}$
std_{IC}: Random ICs on $\boldsymbol{v}$	$2.6\cdot10^{-3}\pm4\cdot10^{-4}$	$1.2\cdot10^{-2}\pm1\cdot10^{-3}$

Table 3.37: (GSS) Trajectory Errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory errors in $\boldsymbol{x}/\boldsymbol{v}$ is calculated using (3.4.3)/(3.4.4).

Since the conservation of the sum of gravitational potential energy and kinetic energy of each planet produces the elliptical orbits around the Sun, we will consider the conservation of total energy (as the sum of gravitational potential energy and kinetic energy) of each planet and Sun as the emergent behavior. The total energy for each planet at time t is calculated as,

$$E_i^{\text{total}}(t) = -\frac{G\tilde{m}_1\tilde{m}_i}{r_{i,1}(t)} + \frac{\tilde{m}_i s_i^2}{2}, \quad \text{for } i = 2,\cdots,5.$$

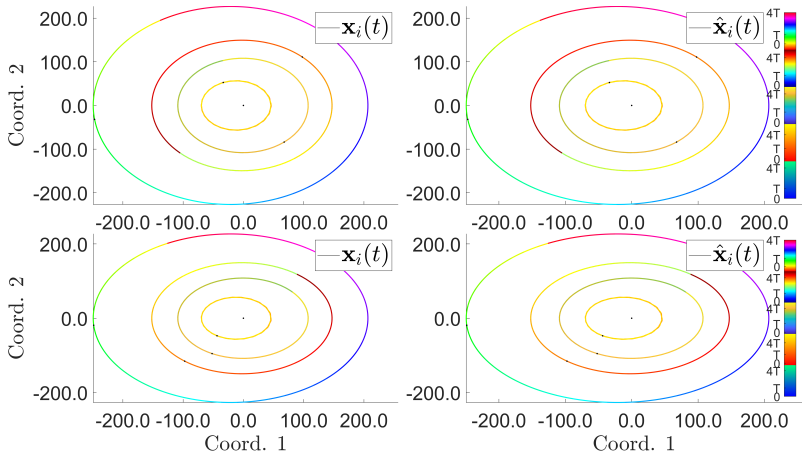


Figure 3.13: (GSS) Comparison of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, with the errors reported in table 3.37. The first row of trajectories are generated from an initial condition taken from the observation data. The second row of trajectories are generated from another randomly chosen initial condition. The first column of trajectories are generated from the true interaction kernel, whereas the second column of trajectories are generated from our estimated kernel with the same initial conditions. The color of the trajectory indicates the flow of time, from $t = 0$ to $t = 4T$; and each AO uses a different set of colors, as given by the color bar on the right.

$r_{i,1}(t) = \|\boldsymbol{x}_i(t) - \boldsymbol{x}_1(t)\|$ and $s_i = \|\boldsymbol{v}_i(t)\|$. Then we consider the variance and mean of the total energy (associated to each planet) over time, i.e.,

$$\begin{cases} E_i^{\text{Mean}} &= \text{Mean}_{l=1}^{L}(E_i^{\text{total}}(t_l)) \\ E_i^{\text{Var}} &= \text{Var}_{l=1}^{L}(E_i^{\text{total}}(t_l)) \end{cases}$$

When $E_i^{\text{Var}} < 10^{-2}$ for $i = 2, \cdots 5$, we consider the total energy to be conserved. Not surprisingly, with the total energy of the true system being always conserved, and with the predicted positions as well as their corresponding velocities of each AO estimated with about 10^{-2} relative errors, 100% of the estimated systems show conservation of total energy. We also consider a set of Pattern Indicator scores to quantitatively measure the capability of our estimators to predict limit cycles correctly for GSS. PI_1 measures the relative errors between the energy variance from the true system and the predicted system over M trials. And PI_2 measures the relative errors between the mean energy from the true system and the predicted system over M trials. The scores

are reported in table 3.38.

	PI_1	PI_2
$mean_{IC}$: Training ICs	$2.85 \cdot 10^{-1} \pm 4 \cdot 10^{-3}$	$4.96 \cdot 10^{-6} \pm 4 \cdot 10^{-8}$
std_{IC}: Training ICs	$1.15 \cdot 10^{-1} \pm 3 \cdot 10^{-3}$	$8.9 \cdot 10^{-7} \pm 2 \cdot 10^{-8}$
$mean_{IC}$: Random ICs	$2.87 \cdot 10^{-1} \pm 5 \cdot 10^{-3}$	$4.94 \cdot 10^{-6} \pm 4 \cdot 10^{-8}$
std_{IC}: Random ICs	$1.16 \cdot 10^{-1} \pm 3 \cdot 10^{-3}$	$9.1 \cdot 10^{-7} \pm 2 \cdot 10^{-8}$

Table 3.38: (GSS) Pattern Indicator Scores: ICs used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows).

Notice that the original system has its total energy variance being close to zero, and we are able to reproduce the total energy variance which is close to zero; moreover, the total energy of the predicted system for each planet resembles closely of its counterpart in the true system.

We have studied the Solar system as an interacting agent-based systems where each agent representing a different type due to the mass-based gravity. However, it is clear that there is only one underlying interaction law with an associated parameter (the mass) for each agent. Our learning approach performs well without any knowledge of this structure and produces each pairwise interaction as a different function. But in fact, they are a family of functions parameterized by one single parameter, which is the mass of each agent. Therefore, in this subsequent section, we proceed to show that the learned functions are close to the known gravitational kernel and that we can discover the underlying masses using an appropriate decoupling procedure.

3.7.1 Discovery of the Parametric Form

Having examined the behaviors of $\hat{\phi}^E_{1,k'}$ and $\hat{\phi}^E_{k',1}$ (for $k' = 2, \cdots, N$) closely, we observe an interesting behavior of our estimators, which is that $\hat{\phi}^E_{1,k'}$ and $\hat{\phi}^E_{k',1}$ (for $k' \neq 1$) behave roughly the same, except at different scales. Such behavior prompts us to

consider a single-parameter parametric structure of $\hat{\phi}^E_{k,k'}$'s, i.e.,

$$\hat{\phi}^E_{k,k'}(r) \approx \beta_{k'}\hat{\phi}^E_m(r) \quad \text{for } k \neq k' \text{ with } \beta_{k'} > 0.$$

Remark 3.7.1. *We do not assume any particular form of $\hat{\phi}^E_m(r)$, except that $\hat{\phi}^E_m(r)$ being continuous.*

In fact, the original gravitational interaction kernels are parameterized by $G\tilde{m}_{i'}$, i.e. $\phi^E_{k,k'}(r) = G\tilde{m}_{k'} \cdot \frac{1}{r^3}$.

Remark 3.7.2. *The gravitational constant G represents the length and time scales on which the experiment is conducted, and it will not be identifiable by our decoupling procedure. Therefore, we assume that G is known. In fact, the first implicit measurement of G with about 1% accuracy is attributed to Henry Cavendish in the Cavendish experiment performed in $1797 - 1798$, and the result was published in Philosophical Transactions of the Royal Society. Using the estimated G, with the radius of Earth first calculated by the Greek mathematician Eratosthenes in approximately 230 BC, and the gravitational acceleration, $g \approx 9.8m/sec^2$, determined by Galileo in the 16^{th} century, one can calculate the mass of the Earth, by connecting Newton's second law and universal law of gravitation, to get $M_{Earth} = 5.98 \cdot 10^{24}kg$.*

Since $\hat{\phi}^E_{k,k'} \approx \phi^E_{k,k'}$ (for $k = 1$ or $k' = 1$), we want to decouple $\beta_{k'}$ and $\hat{\phi}^E_m(r)$ from $\hat{\phi}^E_{k,k'}$ through a three-step optimization procedure. First, we consider a sequence of points $\{r_q\}_{q=1}^Q$ from the supports of $\hat{\phi}^E_{1,k'}$ for $k' = 2, \cdots N$ (r_q's are taken as the centers of the sub-intervals where the basis functions are built), and the following loss function,

$$f_1(\beta_1, \cdots, \beta_N, \hat{\phi}^E_m(r_1), \cdots, \hat{\phi}^E_m(r_Q)) = \sum_{k=2}^{N}\sum_{q=1}^{Q}(\hat{\phi}^E_{k,1}(r_q) - \beta_1\hat{\phi}^E_m(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},k,1}(r_q)$$
$$+ \sum_{k'=2}^{N}\sum_{q=1}^{Q}(\hat{\phi}^E_{1,k'}(r_q) - \beta_{k'}\hat{\phi}^E_m(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},1,k'}(r_q)$$

f_1 is minimized over $\beta_{k'} \geq 0$ for $k' = 1, \cdots, N$ and $\hat{\phi}_m^E(r_q) \in R$ for $q = 1, \cdots, Q$. We only keep portion of the minimizer, namely, $\{\hat{\phi}_m^{\boldsymbol{x},*}(r_q)\}_{q=1}^Q$, due to the fact that the Sun related terms have significantly more dominance in f_1. Second, we extend the discrete values of $\{\hat{\phi}_m^{\boldsymbol{x},*}(r_q)\}_{q=1}^Q$ to a continuous function, and express $\hat{\phi}_m^E$ as a linear combination of basis functions ψ_η (clamped B-spline functions of degree 2) over the interval $[R_1, R_2]$, where

$$R_1 = \min_{k,k'=1,\ldots,K}\{R_{k,k'}^{\min}\}, \quad R_2 = \max_{k,k'=1,\ldots,K}\{R_{k,k'}^{\max}\}, \quad \text{with } \operatorname{supp}(\hat{\phi}_{k,k'}^E) = [R_{k,k'}^{\min}, R_{k,k'}^{\max}].$$

Hence,

$$\hat{\phi}_m^E(r) = \sum_{\eta=1}^{Q} \alpha_\eta \psi_\eta(r).$$

Then, we do a regularized least square fit to $\{\hat{\phi}_m^{\boldsymbol{x},*}(r_q)\}_{q=1}^Q$, using the following loss function,

$$f_2(\alpha_1, \cdots, \alpha_Q) = \sum_{q=1}^{Q}\sum_{\eta=1}^{Q}(\alpha_\eta \psi_\eta(r_q) - \hat{\phi}_m^{\boldsymbol{x},*}(r_q))^2 + \lambda \int_{r=R_1}^{R_2} \Big|\sum_{\eta=1}^{Q} \alpha_\eta \psi_\eta''(r)\Big|^2 dr.$$

Here we take $\lambda = 10^{-3}$. The result of extending the discrete points to a continuous function is shown in Fig. 3.14.

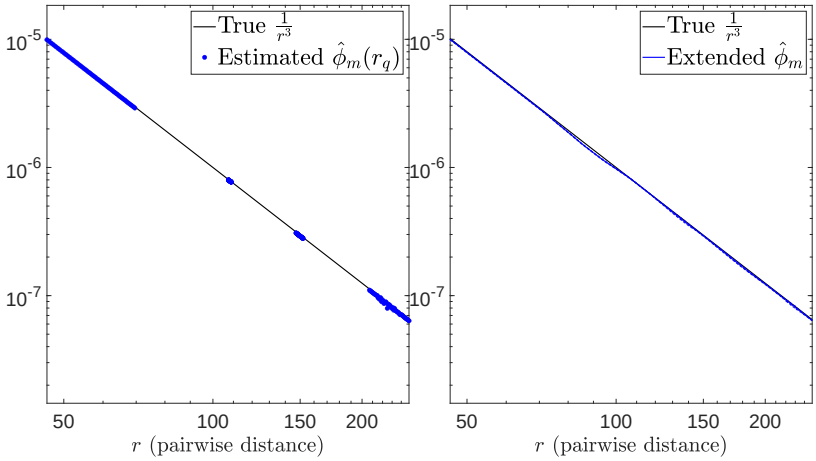


Figure 3.14: (GSS) Extension from discrete $\{\hat{\phi}_m^{\boldsymbol{x},*}(r_q)\}_{q=1}^Q$'s to a continuous $\hat{\phi}_m^E$. The $\frac{1}{r^3}$ line is shown as a reference line and it is not used in the learning of $\{\hat{\phi}_m^{\boldsymbol{x},*}(r_q)\}_{q=1}^Q$'s nor the extension procedure.

The last step is to use the discrete values, $\{\hat{\phi}_m^{\boldsymbol{x},*}(r_q)\}_{q=1}^Q$, to learn the $\beta_{k'}$ again, using the following loss function,

$$f_3(\beta_1,\cdots,\beta_N) = \sum_{k=2}^{N}\sum_{q=1}^{Q} \frac{(\hat{\phi}_{k,1}^E(r_q) - \beta_1 \hat{\phi}_m^{\boldsymbol{x},*}(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},k,1}(r_q)}{\sum_{q=1}^Q (\hat{\phi}_{k,1}^E(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},k,1}(r_q)} + \sum_{k'=2}^{N}\sum_{q=1}^{Q} \frac{(\hat{\phi}_{1,k'}^E(r_q) - \beta_{k'} \hat{\phi}_m^{\boldsymbol{x},*}(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},1,k'}(r_q)}{\sum_{q=1}^Q (\hat{\phi}_{1,k'}^E(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},1,k'}(r_q)}$$

The re-scaling by $\sum_{q=1}^Q (\hat{\phi}_{k,1}^E(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},k,1}(r_q)$ and $\sum_{q=1}^Q (\hat{\phi}_{1,k'}^E(r_q))^2 d\rho_T^{L,M,\boldsymbol{x},1,k'}(r_q)$ is to keep all terms balanced, and in this particular instance it especially counters the dominance of the mass of the Sun, whose mass takes up more than 95% of the mass of the whole solar system. The appropriate use of the dynamics-adapted measures enables us to identify parameters correctly. f_3 is minimized over $\beta_{k'} \geq 0$ for $k' = 1,\cdots,N$. The minimizer $\beta_{k'}^*$ together with $\hat{\phi}_m^E$ (from the previous two steps), will have the following form

$$\beta_{k'} = C_1 \tilde{m}_{k'} \quad \text{for } k' = 1,\cdots,N\,,,$$

and

$$\hat{\phi}_m^E(r) = \frac{C_2}{r^3}, \quad \text{with } C_1 C_2 = G.$$

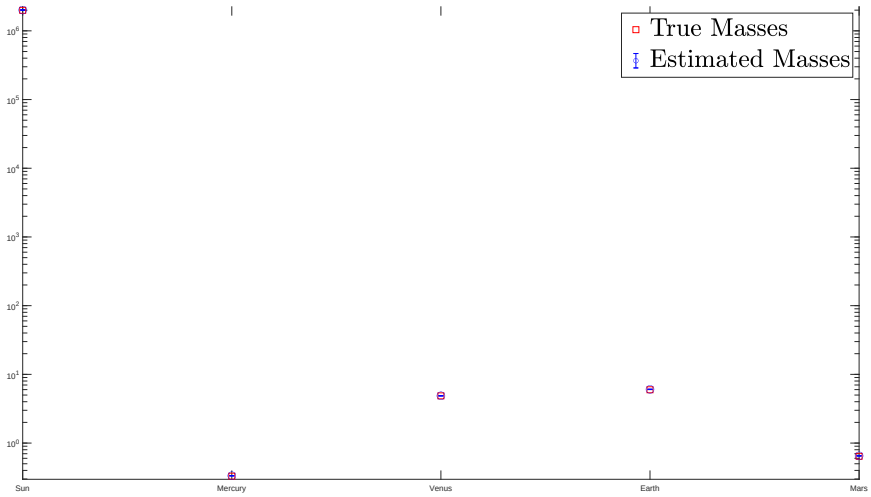


Figure 3.15: (GSS) Comparison of true and mean estimated masses over 10 learning trials.

In order to offer deeper understanding of the difficulty of estimating the masses of each AO, we provide the relative errors for estimating the mass of each astronomical object in table 3.39 along with the mean and standard deviation of the estimated masses.

	Sun	Mercury	Venus	Earth	Mars
True Mass	$1.9885 \cdot 10^6$	$3.3 \cdot 10^{-1}$	4.87	5.97	$6.42 \cdot 10^{-1}$
Estimated Mass	$2.01 \cdot 10^6 \pm 1 \cdot 10^4$	$3.35 \cdot 10^{-1} \pm 2 \cdot 10^{-3}$	$4.88 \pm 3 \cdot 10^{-2}$	$6.05 \pm 4 \cdot 10^{-2}$	$6.52 \cdot 10^{-1} \pm 5 \cdot 10^{-3}$
Rel. Err.	$1.1 \cdot 10^{-2} \pm 7 \cdot 10^{-3}$	$1.4 \cdot 10^{-2} \pm 6 \cdot 10^{-3}$	$4 \cdot 10^{-3} \pm 4 \cdot 10^{-3}$	$1.3 \cdot 10^{-2} \pm 6 \cdot 10^{-3}$	$1.6 \cdot 10^{-2} \pm 8 \cdot 10^{-3}$

Table 3.39: (GSS) True, Estimated Masses, and the Relative Errors. Recall that the masses are measured in unit: 10^{24} kg. Notice the immense difference in the scales of the masses, with the mass of the Sun taking up over 99% of the whole Solar system, which makes the mass estimation problem severely ill-posed.

3.8 Conclusion

We have demonstrated the effectiveness and efficiency of a nonparametric inference procedure to estimate the governing structure of various kinds of collective dynamics from observation of short-time trajectory data. Such estimators can be also used

to predict the correct type of emergent behaviors of the observed systems at larger timescales than those obtained from the training data. The governing models proposed in section 3.2 encompass a wide range of dynamical systems of significant theoretical and computational interests to the physics, biology, and social science communities; and the algorithm in section 3.3 scales efficiently to a large number of homogeneous or heterogeneous agents.

The systems included first-order, one-dimensional interaction kernels (Opinion Dynamics in Sec. 3.5.1), second-order one-dimensional interaction kernels (Cucker-Smale, Self-Propelling Particles in $2D/3D$, in Sec. 3.5.2 to 3.5.4), first-order two-dimensional interaction kernels (Synchronized Oscillator in Sec. 3.6), and second-order families of one-dimensional interaction kernels with underlying, but unknown, single parameters (Gravitational System in Sec. 3.7). In all cases, our estimators exhibit high precision in terms of standard performance measures, as well as high accuracy at capturing the proper type of emergent behaviors as measured by the confusion matrix and pattern indicator scores appropriate to the system. Our final example studied the intrinsic parametric structure of our learned estimators, which leads to the discovery of some fundamental physical concepts, such as accurate mass and the underlying shared kernel of $\frac{1}{r^2}$ for gravitational force.

Further study of more intricate parametric structure of the interaction laws is ongoing as well as the theoretical foundations of the systems (3.2.1),(3.2.2). We are also preparing the study of emergent behaviors on more complex systems with more elaborate interaction laws and governing structures.

Acknolwedgements. We acknowledge support from NSF-ATD-1737984, AFOSR FA9550-17-1-0280, NSF-IIS-1546392, NSF-IIS-1837991, NIH - T32GM119998; we thank Duke University and Prisma Analytics Inc. for free use of computing resources.

3.9 Performance Measures

Similar to what we have defined for measuring the performance of $\widehat{\boldsymbol{\phi}}^E$, we will use $\rho_T^{\xi,k,k'}$ to give the performance indicators of $\widehat{\boldsymbol{\phi}}^\xi$ in first order systems. Similarly we have

$$\begin{cases} \rho_T^{\xi,k,k'}(r,s^\xi) & = \frac{1}{N_{k,k'}T}\int_{t=0}^{T}\mathbb{E}_{\boldsymbol{Y}_0\sim\mu^{\boldsymbol{y}}}\Big[\sum_{\substack{i\in C_k\\ i'\in C_{k'}\\ i\neq i'}}\delta_{r_{i,i'}(t),s^\xi_{i,i'}(t)}(r,s^\xi)\Big]\,dt, \\ \rho_T^{L,\xi,k,k'}(r,s^\xi) & = \frac{1}{N_{k,k'}L}\sum_{l=1}^{L}\mathbb{E}_{\boldsymbol{Y}_0\sim\mu^{\boldsymbol{y}}}\Big[\sum_{\substack{i\in C_k\\ i'\in C_{k'}\\ i\neq i'}}\delta_{r_{i,i'}(t_l),s^\xi_{i,i'}(t_l)}(r,s^\xi)\Big], \\ \rho_T^{L,M,\xi,k,k'}(r,s^\xi) & = \frac{1}{N_{k,k'}LM}\sum_{l,m=1}^{L,M}\sum_{\substack{i\in C_k\\ i'\in C_{k'}\\ i\neq i'}}\delta_{r_{i,i'}(t_l),s^\xi_{i,i'}(t_l)}(r,s^\xi). \end{cases} \tag{3.9.1}$$

For measuring the difference, $\phi^\xi_{k,k'}-\hat{\phi}^\xi_{k,k'}$, we use the following $L^2(\rho_T)$ norm,

$$\left\|\phi^\xi_{k,k'}-\hat{\phi}^\xi_{k,k'}\right\|^2_{L^2(\rho_T^{\xi,k,k'})}=\int_{r=0}^{\infty}\int_{s^\xi=-\infty}^{\infty}(\phi^\xi_{k,k'}(r,s^\xi)-\hat{\phi}^\xi_{k,k'}(r,s^\xi))^2\,d\rho_T^{E,k,k'}(r,s^\xi). \tag{3.9.2}$$

For $(\widehat{\boldsymbol{\phi}}^E,\widehat{\boldsymbol{\phi}}^A,\widehat{\boldsymbol{\phi}}^\xi)$ learned from any second-order system, we will need two new sets of probability distributions. First, for $\rho_T^{E,k,k'}$, it is the same as defined in (3.4.1). Second we define $\rho_T^{\dot{\boldsymbol{x}},k,k'}$ as follows,

$$\begin{cases} \rho_T^{\dot{\boldsymbol{x}},k,k'}(r,s^{\dot{\boldsymbol{x}}},\dot{r}) & = \frac{1}{N_{k,k'}T}\int_{t=0}^{T}\mathbb{E}_{\boldsymbol{Y}_0\sim\mu^{\boldsymbol{y}}}\Big[\sum_{\substack{i\in C_k\\ i'\in C_{k'}\\ i\neq i'}}\delta_{r_{i,i'}(t),s^{\dot{\boldsymbol{x}}}_{i,i'}(t),\dot{r}_{i,i'}(t)}(r,s^{\dot{\boldsymbol{x}}},\dot{r})\Big]\,dt, \\ \rho_T^{L,\dot{\boldsymbol{x}},k,k'}(r,s^{\dot{\boldsymbol{x}}},\dot{r}) & = \frac{1}{N_{k,k'}L}\sum_{l=1}^{L}\mathbb{E}_{\boldsymbol{Y}_0\sim\mu^{\boldsymbol{y}}}\Big[\sum_{\substack{i\in C_k\\ i'\in C_{k'}\\ i\neq i'}}\delta_{r_{i,i'}(t_l),s^{\dot{\boldsymbol{x}}}_{i,i'}(t_l),\dot{r}_{i,i'}(t_l)}(r,s^{\dot{\boldsymbol{x}}},\dot{r})\Big], \\ \rho_T^{L,M,\dot{\boldsymbol{x}},k,k'}(r,s^{\dot{\boldsymbol{x}}},\dot{r}) & = \frac{1}{N_{k,k'}LM}\sum_{l,m=1}^{L,M}\sum_{\substack{i\in C_k\\ i'\in C_{k'}\\ i\neq i'}}\delta_{r_{i,i'}(t_l),s^{\dot{\boldsymbol{x}}}_{i,i'}(t_l),\dot{r}_{i,i'}(t_l)}(r,s^{\dot{\boldsymbol{x}}},\dot{r}). \end{cases} \tag{3.9.3}$$

Here, $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{V} \\ \boldsymbol{\Xi} \end{bmatrix}$, $\mu^{\boldsymbol{y}} = \begin{bmatrix} \mu^{\boldsymbol{x}} \\ \mu^{\dot{\boldsymbol{x}}} \\ \mu^{\xi} \end{bmatrix}$, and $\dot{r}$, being not the derivative r, rather the pairwise speed data, e.g., $\dot{r}_{i,i'}(t) = \|\boldsymbol{v}_{i'}(t) - \boldsymbol{v}_i(t)\|$. Finally, $\rho_T^{\xi,k,k'}$ is defined slightly differently from (3.9.1),

$$\begin{cases} \rho_T^{\xi,k,k'}(r, s^\xi, \xi) &= \frac{1}{N_{k,k'}T} \int_{t=0}^T \mathbb{E}_{\boldsymbol{Y}_0 \sim \mu^{\boldsymbol{y}}} \Big[\sum_{\substack{i \in C_k \\ i' \in C_{k'} \\ i \neq i'}} \delta_{r_{i,i'}(t), s^\xi_{i,i'}(t), \xi_{i,i'}(t)}(r, s^\xi, \xi) \Big] \, dt, \\ \rho_T^{L,\xi,k,k'}(r, s^\xi, \xi) &= \frac{1}{N_{k,k'}L} \sum_{l=1}^L \mathbb{E}_{\boldsymbol{Y}_0 \sim \mu^{\boldsymbol{y}}} \Big[\sum_{\substack{i \in C_k \\ i' \in C_{k'} \\ i \neq i'}} \delta_{r_{i,i'}(t_l), s^\xi_{i,i'}(t_l), \xi_{i,i'}(t)}(r, s^\xi, \xi) \Big], \\ \rho_T^{L,M,\xi,k,k'}(r, s^\xi) &= \frac{1}{N_{k,k'}LM} \sum_{l,m=1}^{L,M} \sum_{\substack{i \in C_k \\ i' \in C_{k'} \\ i \neq i'}} \delta_{r_{i,i'}(t_l), s^\xi_{i,i'}(t_l)}(r, s^\xi). \end{cases} \tag{3.9.4}$$

Here, $\xi_{i,i'}(t) = \big|\xi_{i'}(t) - \xi_i(t)\big|$. The prediction error, $\phi^E_{k,k'} - \hat{\phi}^E_{k,k'}$, is measured in the same norm defined in (3.4.2); for $\phi^\xi_{k,k'} - \hat{\phi}^\xi_{k,k'}$, but it is weighted differently,

$$\Big\| \phi^\xi_{k,k'} - \hat{\phi}^\xi_{k,k'} \Big\|^2_{L^2(\rho_T^{\xi,k,k'})} = \int_{r=0}^\infty \int_{s^\xi=-\infty}^\infty (\phi^\xi_{k,k'}(r, s^\xi) - \hat{\phi}^\xi_{k,k'}(r, s^\xi))^2 \, \xi^2 d\rho_T^{E,k,k'}(r, s^\xi, \xi). \tag{3.9.5}$$

and for $\phi^A_{k,k'} - \hat{\phi}^A_{k,k'}$, the corresponding norm is defined as follows,

$$\Big\| \phi^A_{k,k'} - \hat{\phi}^A_{k,k'} \Big\|^2_{L^2(\rho_T^{\dot{\boldsymbol{x}},k,k'})} = \int_{r=0}^\infty \int_{s^{\dot{\boldsymbol{x}}}=-\infty}^\infty (\phi^A_{k,k'}(r, s^{\dot{\boldsymbol{x}}}) - \hat{\phi}^A_{k,k'}(r, s^{\dot{\boldsymbol{x}}}))^2 \, \dot{r}^2 d\rho_T^{\dot{\boldsymbol{x}},k,k'}(r, s^{\dot{\boldsymbol{x}}}, \dot{r}). \tag{3.9.6}$$

Chapter 4

Extension to Manifolds

4.1 Introduction

Let $(\mathcal{M}, g)$ be a connected, smooth, and geodesically-complete d-dimensional Riemannian manifold, with the Riemannian distance denoted by $d_{\mathcal{M}}$. Consider N interacting agents, each represented by a state vector $\boldsymbol{x}_i(t) \in \mathcal{M}$. Their dynamics is governed by the following first order dynamical system, where ϕ, the *interaction kernel*, is the object of our inference: for each $i = 1, \ldots, N$,

$$\dot{\boldsymbol{x}}_i(t) = \frac{1}{N} \sum_{i'=1}^{N} \phi(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))) \boldsymbol{w}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)) \,. \tag{4.1.1}$$

Here $\boldsymbol{w}(\boldsymbol{z}_1, \boldsymbol{z}_2)$, for $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}$, is a weight vector pointing in the tangent direction at $\boldsymbol{z}_1$ to the shortest geodesic from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$. For this to make sense, we restrict our attention to local interactions, e.g. by assuming that ϕ is compactly supported in a sufficiently small interval $[0, R]$, so that length-minimizing geodesics exist uniquely. We discuss the well-posedness of this model in greater detail in section 4.2.1, where we emphasize that this model is derived naturally as a gradient system with a special potential energy depending on pairwise Riemannian distances.

Our observations consist of states along multiple trajectories, namely $\{\boldsymbol{x}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$

with L being the number of observations made in time and M being the number of trajectories. We construct an estimator $\hat{\phi}_{L,M,\mathcal{H}}$ of ϕ that is both close to ϕ in an appropriate L^2 sense, and generates a system in the form of (4.1.1) with accurate trajectories when compared to the observed trajectories (generated by ϕ) with the same initial condition. The estimator, $\hat{\phi}_{L,M,\mathcal{H}}$, is defined as the solution to the minimization problem

$$\hat{\phi}_{L,M,\mathcal{H}} = \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,M,\mathcal{M}}(\varphi)$$

Here $\mathcal{H}$ is a special function space containing suitable approximations to ϕ and $\mathcal{E}_{L,M,\mathcal{M}}$ is a least squares loss functional built from the trajectory data, which also takes into account the underlying geometry of $(\mathcal{M}, g)$. Having established a geometry-based coercivity condition that ensures, among other things, the recoverability of ϕ by a suitable sequence of $\hat{\phi}_{L,M,\mathcal{H}}$'s, our theory shows the convergence rate (in M) of our estimator to the true interaction kernel is independent of the dimension of the observation data, i.e. Nd, and is the same as the minimax rate for 1-dimensional nonparametric regression:

$$\mathbb{E}_{\boldsymbol{X}_0 \sim \mu^{\boldsymbol{x}}} \Big[\Big\| \hat{\phi}_{L,M,\mathcal{H}}(\cdot) \cdot - \phi(\cdot) \cdot \Big\|_{L^2(\rho^L_{T,\mathcal{M}})} \Big] \leq C_1(\mathcal{M}) \left(\frac{\log M}{M} \right)^{\frac{s}{2s+1}}.$$

Here $\boldsymbol{X}_0 \in \mathcal{M}^N$ is an initial system state, $\mu^{\boldsymbol{x}}$ is a distribution of initial system states on $\mathcal{M}^N$, $\rho^L_{T,\mathcal{M}}$ is a dynamics-adapted probability measure which captures the distribution of pairwise Riemannian distances, and $C_1(\mathcal{M})$ is a constant depending the geometry of $\mathcal{M}$ (see sec. 4.4.3).

We also establish bounds on the error between the trajectories evolved using our estimators and the true trajectories. Let $\hat{\boldsymbol{X}}_{[0,T]}, \boldsymbol{X}_{[0,T]}$ be trajectories evolved with the interaction kernels $\hat{\phi}_{L,M,\mathcal{H}}$ and ϕ respectively, started at the same initial condition,

then:

$$\mathbb{E}_{\boldsymbol{X}_0\sim\mu^{\boldsymbol{x}}}\left[d_{\mathrm{trj}}(\boldsymbol{X}_{[0,T]},\hat{\boldsymbol{X}}_{[0,T]})^2\right] \leq C_2(\mathcal{M})\left\|\phi(\cdot)\cdot - \hat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot\right\|^2_{L^2(\rho_{T,\mathcal{M}})},$$

where d_{trj} is a natural geometry-based distance on trajectories and $C_2(\mathcal{M})$ is a constant depending on the manifold's geometry. As M grows, the norm on the right hand side converges at the rate above, yielding convergence of the trajectories; full details are given in section 4.4.4.

The numerical details of the algorithms for learning the estimator and computing trajectories on manifolds are presented in the Appendix. The essential differences, compared to the algorithms presented for Euclidean spaces, are the use of a geometric numerical integrator for computing the evolution of the manifold-constrained dynamics, and that at every time step we need to compute Riemannian inner products of tangent vectors, geodesics and Riemannian distances. We demonstrate the performances of our estimators on an opinion dynamics and a predator-swarm model, each constrained on two model spaces: the two dimensional sphere $\mathbb{S}^2$ and the Poincaré disk.

4.1.1 Connections and Related Work

The research on inferring a suitable dynamical system of interacting agents from observation data has been a longstanding problem in science and engineering; see [91, 74, 50, 130] and references therein. Many recent approaches in machine learning have been developed for inferring general dynamical systems, including multistep methods [75], optimization [141], sparse regression [24, 115, 118], Bayesian regression [145], and deep learning [109, 114]. In a different direction, the generalization of traditional machine learning algorithms in Euclidean settings to Riemannian manifolds, and the development of new algorithms designed to work on Riemannian manifolds, has been attracting increasing attention; for example in variational calculus [124],

reinforcement learning [112], deep learning [31] and theoretical CS [97].

4.2 Model Equations

In this section we introduce the governing equations which we use to model interacting agents constrained on Riemannian manifolds, and discuss the properties of the dynamics. Table 4.1 shows a list of definitions of the common terms used throughout this chapter.

Variable	Definition
$(\mathcal{M}, g)$	Riemannian Manifold with metric g
$T_{\boldsymbol{x}}\mathcal{M}$	Tangent plane to $\mathcal{M}$ at $\boldsymbol{x}$
$\langle \cdot, \cdot, \rangle\, g(\boldsymbol{x})$, $\langle \cdot, \cdot, \rangle\, g$	Inner product on $T_{\boldsymbol{x}}\mathcal{M}$
$\|\boldsymbol{v}\|_{T_{\boldsymbol{x}}\mathcal{M}}$, $\|\boldsymbol{v}\|_g$	Length of $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ induced by $g(\boldsymbol{x})$
$d_{\mathcal{M}}(\cdot, \cdot)$	Riemannian distance induced by g
$C^1(\mathcal{X})$	Set of cont. diff. functions on $\mathcal{X} \subset \mathbb{R}^d$

Table 4.1: Notation for first-order models, also see the Appendix.

4.2.1 Main model

In order to motivate the choice of the model equations we use, we begin with a geometric gradient flow model of an interacting agent system. Consider a system of N interacting agents, with each agent described by a state vector $\boldsymbol{x}_i(t)$ on a d-dimensional connected, smooth, and geodesically complete Riemannian manifold $\mathcal{M}$ with metric g. The change of the state vectors seeks to decrease a system energy E:

$$\frac{d\boldsymbol{x}_i(t)}{dt} = -\partial_{\boldsymbol{x}_i} E(\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_N(t)), \qquad i = 1, \dots, N.$$

Our first key assumption is that E takes the special form

$$E(\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_N(t)) = \frac{1}{N} \sum_{i'=1}^{N} U(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))^2),$$

for some $U : \mathbb{R}^+ \to \mathbb{R}$ with $U(0) = 0$, and $d_{\mathcal{M}}(\cdot, \cdot)$ the geodesic distance on $(\mathcal{M}, g)$. Simplifying, and omitting from the notation the dependency on t of $\dot{\boldsymbol{x}}_i$ and $\boldsymbol{x}_i$, we obtain the first-order geometric evolution equation,

$$\dot{\boldsymbol{x}}_i = \frac{1}{N} \sum_{i'=1}^{N} \phi(d_{\mathcal{M}}(\boldsymbol{x}_i, \boldsymbol{x}_{i'})) \boldsymbol{w}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}), \tag{4.2.1}$$

for $i = 1, \ldots, N$. We call $\phi(r) \coloneqq 2U'(r^2)$ the *interaction kernel*. We have let $\boldsymbol{w}(\boldsymbol{z}_1, \boldsymbol{z}_2) \coloneqq d_{\mathcal{M}}(\boldsymbol{z}_1, \boldsymbol{z}_2) \boldsymbol{v}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ for $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}$, with $\boldsymbol{v}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ being, for $\boldsymbol{z}_2 \neq \boldsymbol{z}_1$, the unit vector (i.e. $\|\boldsymbol{v}\|_{T_{\boldsymbol{z}_1}\mathcal{M}} = 1$) tangent at $\boldsymbol{z}_1$ to the minimizing geodesic from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$ if $\boldsymbol{z}_2$ not in the cut locus of $\boldsymbol{z}_1$, and equal to $\mathbf{0}$ otherwise. In order to guarantee existence and uniqueness of a solution for (4.2.1) over the time interval $[0, T]$, we make a further assumption that ϕ belongs to the admissible space

$$\mathcal{K}_{R,S} \coloneqq \{\varphi \in C^1([0, R]) \Big| \, \|\varphi\|_{L^\infty} + \|\varphi'\|_{L^\infty} \leq S\},$$

for some constant $S > 0$. Here, R is smaller than the global injectivity radius of $\mathcal{M}$, and $L^\infty = L^\infty([0, R])$. With this assumption, the possible discountinuity of $\boldsymbol{v}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ due to either $\boldsymbol{z}_2 \to \boldsymbol{z}_1$ or $\boldsymbol{z}_2$ tends to a point in the cut locus of $\boldsymbol{z}_1$ is canceled by the multiplication by $d_{\mathcal{M}}(\boldsymbol{z}_1, \boldsymbol{z}_2) \to 0$ in the former case, and $\phi(d_{\mathcal{M}}(\boldsymbol{z}_1, \boldsymbol{z}_2)) \to 0$ in the latter case. Therefore, the ODE system in (4.2.1) has a Lipschitz right hand side, thus it has a unique solution existing for $t \in [0, T]$ [67].

With this geometric gradient flow point of view, the from of the equations and the radial symmetry of the interaction kernels are naturally pre-determined by the energy potential. This approach seems to us natural and geometric; for different approaches see [6, 26]. Note that in the case of $\mathcal{M} = \mathbb{R}^d$ with the Euclidean metric, we have $d_{\mathcal{M}}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|$ and $\boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \frac{\boldsymbol{x}_{i'} - \boldsymbol{x}_i}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|}$, and we recover the Euclidean space models used in [19, 89] and the many works referenced therein.

4.3 Learning Framework

We are given a set of trajectory data of the form $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$, for $0 = t_1 < \ldots < t_L = T$, with the initial conditions $\{\boldsymbol{x}_i^m(0)\}_{i=1}^N$ being i.i.d from a distribution $\mu_0(\mathcal{M})$. The objective is to construct an estimator $\hat{\phi}_{L,M,\mathcal{H}}$ of the interaction kernel ϕ.

Before we describe the construction of our estimator, we introduce some notation. We let, in $\mathcal{M}^N \coloneqq \mathcal{M} \times \cdots \times \mathcal{M}$,

$$\boldsymbol{X}_{t_l}^m \coloneqq \begin{bmatrix} \vdots \\ \boldsymbol{x}_i^m(t_l) \\ \vdots \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} \coloneqq \begin{bmatrix} \vdots \\ \boldsymbol{x}_i \\ \vdots \end{bmatrix},$$

where $(\mathcal{M}^N, g_{\mathcal{M}}^N)$ is the canonical product of Riemannian manifolds with product Riemannian metric given by,

$$\left\langle \begin{bmatrix} \vdots \\ \boldsymbol{u}_i \\ \vdots \end{bmatrix}, \begin{bmatrix} \vdots \\ \boldsymbol{z}_i \\ \vdots \end{bmatrix} \right\rangle_{g_{\mathcal{M}}^N(\boldsymbol{X})} \coloneqq \frac{1}{N} \sum_{i=1}^N \langle \boldsymbol{u}_i, \boldsymbol{z}_{i,\rangle}\, g(\boldsymbol{x}_i),$$

for $\boldsymbol{u}_i, \boldsymbol{z}_i \in T_{\boldsymbol{x}_i}\mathcal{M}$. The initial conditions, $\boldsymbol{X}_0^m$ are drawn i.i.d. from $\mu^{\boldsymbol{x}}$, where $\mu^{\boldsymbol{x}} = \mu_0(\mathcal{M}) \times \cdots \times \mu_0(\mathcal{M}) = \mu_0(\mathcal{M})^N$. Note that all expectations will be with respect to $\boldsymbol{X}_0 \sim \mu^{\boldsymbol{x}}$. Finally, $\boldsymbol{f}_\phi^{\mathrm{c}}$ is the vector field on $\mathcal{M}^N$ (i.e. $\boldsymbol{f}_\phi^{\mathrm{c}}(\boldsymbol{X}) \in T_{\boldsymbol{X}}\mathcal{M}^N$ for $\boldsymbol{X} \in \mathcal{M}^N$), given by

$$\boldsymbol{f}_\phi^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m) \coloneqq \begin{bmatrix} \vdots \\ \frac{1}{N}\sum_{i'=1}^N \phi(d_{\mathcal{M}}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l)))\boldsymbol{w}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l)) \\ \vdots \end{bmatrix},$$

The system of equations (4.2.1) can then be rewritten, for each $m = 1, \ldots, M$, as

$$\dot{\boldsymbol{X}}_t^m = \boldsymbol{f}_\phi^{\mathrm{c}}(\boldsymbol{X}_t^m)\,.$$

4.3.1 Geometric Loss Functionals

In order to simplify the presentation, we assume that the observation times, i.e. $\{t_l\}_{l=1}^L$, are equispaced in $[0,T]$ (the general case is similar). We begin with the definition of the hypothesis space $\mathcal{H}$, over which we shall minimize an error functional to obtain an estimator of ϕ.

Definition 4.3.1. *An admissible hypothesis space $\mathcal{H}$ is a compact (in L^∞-norm) and convex subset of $L^2([0,R])$, such that every $\varphi \in \mathcal{H}$ is bounded above by some constant $S_0 \geq S$, i.e. $\|\varphi\|_{L^\infty([0,R])} \leq S_0$; moreover φ is smooth enough to ensure the existence and uniqueness of solutions of (4.2.1) for $t \in [0,T]$, i.e. $\varphi \in \mathcal{H} \cap \mathcal{K}_{R,S_0}$.*

For a function $\varphi \in \mathcal{H}$, we define the loss functional

$$\mathcal{E}_{L,M,\mathcal{M}}(\varphi) \coloneqq \frac{1}{ML} \sum_{l,m=1}^{L,M} \left\| \dot{\boldsymbol{X}}_{t_l}^m - \boldsymbol{f}_\varphi^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m) \right\|_g^2 , \tag{4.3.1}$$

where the norm $\|\cdot\|_g$ in $T_{\boldsymbol{X}_{t_l}^m}\mathcal{M}^N$ can be written as

$$\left\| \dot{\boldsymbol{X}}_{t_l}^m - \boldsymbol{f}_\varphi^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m) \right\|_g^2 = \frac{1}{N} \sum_{i=1}^N \left\| \dot{\boldsymbol{x}}_{i,t_l}^m - \frac{1}{N} \sum_{i'=1}^N \varphi(r_{ii',t_l}^m) \boldsymbol{w}_{ii',t_l}^m \right\|_{T_{\boldsymbol{x}_i^m(t_l)}\mathcal{M}}^2 ,$$

with $\dot{\boldsymbol{x}}_{i,t_l}^m \coloneqq \dot{\boldsymbol{x}}_i^m(t_l)$, $r_{ii',t_l}^m \coloneqq d_{\mathcal{M}}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l))$, and $\boldsymbol{w}_{ii',t_l}^m \coloneqq \boldsymbol{w}(\boldsymbol{x}_i^m(t_l), \boldsymbol{x}_{i'}^m(t_l))$. This loss functional is nonnegative, and reaches 0 when φ is equal to the (true) interaction kernel ϕ if ϕ is also in $\mathcal{H}$ (i.e. $\phi \in \mathcal{H} \cap \mathcal{K}_{R,S}$). Given that $\mathcal{H}$ is compact and convex, $\mathcal{E}_{L,M,\mathcal{M}}$ is continuous on $\mathcal{H}$, the minimizer of $\mathcal{E}_{L,M,\mathcal{M}}$ exists and is unique. We define it to be out estimator:

$$\hat{\phi}_{L,M,\mathcal{H}} \coloneqq \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,M,\mathcal{M}}(\varphi) .$$

As $M \to \infty$, by the law of large numbers, we have $\mathcal{E}_{L,M,\mathcal{M}} \to \mathcal{E}_{L,\infty,\mathcal{M}}$, with

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) \coloneqq \frac{1}{L} \sum_{l=1}^L \mathbb{E}\Big[\left\| \dot{\boldsymbol{X}}_{t_l} - \boldsymbol{f}_\varphi^{\mathrm{c}}(\boldsymbol{X}_{t_l}) \right\|_g^2 \Big]. \tag{4.3.2}$$

Since $\mathcal{E}_{L,\infty,\mathcal{M}}$ is continuous on $\mathcal{H}$, the minimization of $\mathcal{E}_{L,\infty,\mathcal{M}}$ over $\mathcal{H}$ is well-posed and it has a unique minimizer $\hat{\phi}_{L,\infty,\mathcal{H}} \coloneqq \arg\min_{\varphi\in\mathcal{H}} \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi)$. Much of our theoretical work establishes the relationship between the estimator $\hat{\phi}_{L,M,\mathcal{H}}$, the closely related (in the infinite sample limit $M \to \infty$) $\hat{\phi}_{L,\infty,\mathcal{H}}$, and the true interaction kernel ϕ.

4.3.2 Performance Measures

We introduce a suitable normed function space in which to compare the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ with the true interaction kernel ϕ. We also measure performance in terms of trajectory estimation error based on a distance between trajectories generated from the true dynamics (evolved using ϕ with some initial condition $\boldsymbol{X}_0 \sim \mu^{\boldsymbol{x}}$) and the estimated dynamics (evolved using the estimated interaction kernel $\hat{\phi}_{L,M,\mathcal{H}}$, and with the same initial condition, i.e. $\boldsymbol{X}_0$).

Estimation Error

First we introduce a probability measure $\rho_{T,\mathcal{M}}$ on $\mathbb{R}_+$, that is used to define a norm to measure the error of the estimator, derived from the loss functionals (given by (4.3.1) and (4.3.2)), that reflects the distribution of pairwise data given by the dynamics as well as the geometry of the manifold $\mathcal{M}$:

$$\rho_{T,\mathcal{M}}(r) \coloneqq \frac{1}{\binom{N}{2}}\mathbb{E}\Big[\frac{1}{T}\int_0^T \sum_{i,i'} \delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t))}(r)\,dt\Big],$$

where δ is the Dirac delta function. In words, this measure is obtained by averaging δ-functions having mass at any pairwise distances in any trajectory, over all initial conditions drawn from $\mu^{\boldsymbol{x}}$, over all pairs of agents and all times. A time-discretized version is given by:

$$\rho^L_{T,\mathcal{M}}(r) \coloneqq \frac{1}{L\binom{N}{2}}\mathbb{E}\Big[\sum_{l=1}^{L}\sum_{1\le i<i'\le N} \delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t_l),\boldsymbol{x}_{i'}(t_l))}(r)\Big].$$

The two probability measures defined above appear naturally in the proofs for the convergence rate of the estimator. From observational data we compute the empirical version:

$$\rho_{T,\mathcal{M}}^{L,M}(r) \coloneqq \frac{1}{ML\binom{N}{2}} \sum_{l,m=1}^{L,M} \sum_{1\le i<i'\le N} \delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t_l),\boldsymbol{x}_{i'}(t_l))}(r).$$

The geometry of $\mathcal{M}$ is incorporated in these three measures by the presence of geodesic distances. The norm

$$\|\varphi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}^2 \coloneqq \int_{r=0}^{\infty} \left|\varphi(r)r\right|^2 d\rho_{T,\mathcal{M}}(r)$$

is used to define the estimation error: $||\hat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot||_{L^2(\rho_{T,\mathcal{M}})}$. We also use a relative version of this error, to enable a meaningful comparison across different interaction kernels:

$$\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{\mathrm{Rel}.L^2(\rho_{T,\mathcal{M}})} \coloneqq \frac{\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}{\|\phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}. \tag{4.3.3}$$

Trajectory Estimation Error

Let $\boldsymbol{X}_{[0,T]}^m \coloneqq (\boldsymbol{X}_t^m)_{t\in[0,T]}$ be the trajectory generated by the m^{th} initial condition, $\boldsymbol{X}_0^m$. The trajectory estimation error between $\boldsymbol{X}_{[0,T]}^m$ and $\hat{\boldsymbol{X}}_{[0,T]}^m$, evolved using, the unknown interaction kernel ϕ and, respectively, the estimated one, $\hat{\phi}$, with the same initial condition, is given by

$$d_{\mathrm{trj}}(\boldsymbol{X}_{[0,T]}^m, \hat{\boldsymbol{X}}_{[0,T]}^m)^2 \coloneqq \sup_{t\in[0,T]} \frac{\sum_i d_{\mathcal{M}}(\boldsymbol{x}_i^m(t), \hat{\boldsymbol{x}}_i^m(t))^2}{N}. \tag{4.3.4}$$

We are also interested in the performance over different initial conditions, hence we use $\mathrm{mean}_{\mathrm{IC}}$ and $\mathrm{std}_{\mathrm{IC}}$ to report the mean and std of these trajectory errors over a (large) number of initial conditions sampled i.i.d. from $\mu^{\boldsymbol{x}}$.

4.3.3 Algorithm

The algorithm in section 4.9 shows the detailed steps on how to construct the estimator to ϕ given the observation data.

4.3.4 Computational Complexity

Assuming a finite dimensional subspace of $\mathcal{H}$, i.e. $\mathcal{H}_M \subset \mathcal{H}$ with $d(\mathcal{H}_M) = n(M)$, we are able to re-write the minimization problem of (4.3.1) over $\mathcal{H}_M$ as a linear system, i.e. $A_M \vec{\alpha} = \vec{b}_M$ with $A_M \in R^{n\times n}$ and $\vec{b}_M \in \mathbb{R}^{n\times 1}$; for details, see the Appendix. This linear system is well conditioned, ensured by the geometric coercivity condition.

The total computational cost for solving the learning problem is: $MLN^2 + MLdn^2 + n^3$ with MLN^2 for computing pairwise distances, $MLdn^2$ for assembling A_M and $\vec{b}_M$, and n^3 for solving $A_M\vec{\alpha} = \vec{b}_M$. When choosing the optimal $n = n_* \approx (\frac{M}{\log M})^{\frac{1}{2s+1}} \approx M^{\frac{1}{3}}$ ($s = 1$ for C^1 functions) as per Thm. 4.8.6, we have comp. time $= MLN^2 + MLdM^{\frac{2}{3}} + M = \mathcal{O}(M^{\frac{5}{3}})$. The computational bottleneck comes from the assembly of A_M and $\vec{b}_M$. However, since we can parallelize our learning approach in m, the updated computing time in the parallel regime is comp. time $= \mathcal{O}\Big(\Big(\frac{M}{\text{num. cores}}\Big)^{\frac{5}{3}}\Big)$. The total storage for the algorithm is $MLNd$ floating-point numbers for the trajectory data, albeit one does not need to hold all of the trajectory data in memory. The algorithm can process the data from one trajectory at a time, requiring LNd. Once the linear system, $A_M\vec{\alpha} = \vec{b}_M$, is assembled, the algorithm just needs to hold roughly n^2 floating-point numbers in memory. When we use the optimal number of basis functions, i.e. $n_* = M^{\frac{1}{3}}$, the memory used is $O(M^{\frac{2}{3}})$.

4.4 Learning Theory

We present in this section the major results establishing the convergence of the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ to ϕ, at the optimal learning rate, and bounding the trajectory

estimation error between the true and estimated dynamics (evolved using $\hat{\phi}_{L,M,\mathcal{H}}$), with their corresponding proofs in the Appendix.

4.4.1 Learnability: geometric coercivity condition

We establish a geometry-adapted coercivity condition, extending that of [19, 89] to the Riemannian setting, which will guarantee the uniqueness of the minimizer of $\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi)$, and that $\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi)$ controls the $\|\cdot\|_{L^2(\rho_{T,\mathcal{M}})}$ distance between the minimizer and the true interaction kernel.

Definition 4.4.1 (Geometric Coercivity condition)**.** *The geometric evolution system in* (4.2.1) *with initial condition sampled from $\mu^{\boldsymbol{x}}$ on $\mathcal{M}^N$ is said to satisfy the geometric coercivity condition on the admissible hypothesis space $\mathcal{H}$ if there exists a constant $c \equiv c_{L,N,\mathcal{H},\mathcal{M}} > 0$ such that for any $\varphi \in \mathcal{H}$ with $\varphi(\cdot)\cdot \in L^2(\rho^L_{T,\mathcal{M}})$ we have*

$$c\,\|\varphi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\Big[\,\big\|\boldsymbol{f}^c_\varphi(\boldsymbol{X}_{t_l})\big\|^2_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}\Big].$$

In order to simplify the argument on how this geometric coercivity condition controls the distance between $\hat{\phi}_{L,\infty,\mathcal{H}}$ and ϕ, we introduce an inner product on $L^2 = L^2(\rho^L_{T,\mathcal{M}})$:

$$\langle\langle\varphi_1,\varphi_2\rangle\rangle_{L^2} \coloneqq \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\Big[\langle \boldsymbol{f}^{\mathrm{c}}_{\varphi_1}(\boldsymbol{X}_{t_l}), \boldsymbol{f}^{\mathrm{c}}_{\varphi_2}(\boldsymbol{X}_{t_l})_{,\rangle}\, T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N\Big].$$

Then the geometric coercivity condition can be rewritten as

$$c_{L,N,\mathcal{H},\mathcal{M}}\,\|\varphi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \langle\langle\varphi,\varphi\rangle\rangle_{L^2(\rho^L_{T,\mathcal{M}})},$$

and since the loss function from (4.3.2) can be written as $\mathcal{E}_{L,\infty,\mathcal{H}}(\varphi) = \langle\langle\varphi-\phi,\varphi-\phi\rangle\rangle$, this implies

$$c_{L,N,\mathcal{H},\mathcal{M}}\,\|\varphi(\cdot)\cdot -\phi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \mathcal{E}_{L,\infty,\mathcal{H}}(\varphi).$$

Hence when $\mathcal{E}_{L,\infty,\mathcal{H}}(\varphi)$ is small, $\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}$ is also small; hence if we construct a sequence of minimizers of $\mathcal{E}_{L,\infty,\mathcal{H}}$ over increasing $\mathcal{H}$ with decreasing $\mathcal{E}_{L,\infty,\mathcal{H}}$ values, the convergence of $\hat{\phi}_{L,\infty,\mathcal{H}}$ to ϕ can be established.

4.4.2 Concentration and Consistency

The first theorem bounds, with high probability, the difference between the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ and the true interaction kernel ϕ, which makes apparent the trade-off between the $L^2(\rho^L_{T,\mathcal{M}})$-distance between ϕ and $\mathcal{H}$ (approximation error), and M the number of trajectories needed for achieving the desired accuracy. Here $\mathcal{N}(\mathcal{U},\epsilon)$ is the covering number of a set $\mathcal{U}$ with open balls of radius ϵ w.r.t the L^∞-norm.

Theorem 4.4.1. *Let $\phi \in L^2([0,R])$, and $\mathcal{H}$ an admissible hypothesis space such that the geometric coercivity condition holds with a constant $c_{L,N,\mathcal{H},\mathcal{M}}$. Then, $\hat{\phi}_{L,M,\mathcal{H}}$, minimizer of* (4.3.1) *on the trajectory data generated by* (4.2.1)*, satisfies*

$$\left\|\hat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\right\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \frac{2}{c_{L,N,\mathcal{H},\mathcal{M}}}\Big(\epsilon + \inf_{\varphi\in\mathcal{H}} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})}\Big)$$

with probability at least $1-\tau$, when $M \geq \frac{1152 S_0^2 R^2}{\epsilon c_{L,N,\mathcal{H},\mathcal{M}}}(\ln \mathcal{N}(\mathcal{H}, \frac{\epsilon}{48 S_0 R^2}) + \ln\frac{1}{\tau})$.

This quantifies the usual bias-variance tradeoff in our setting: on the one hand, with a large hypothesis space, the quantity $\inf_{\varphi\in\mathcal{H}} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}$ could be made small. On the other hand, we wish to have the right number of samples to make the variance of the estimator small, by controlling the covering number of the hypothesis space $\mathcal{H}$.

4.4.3 Convergence Rate

Next we establish the convergence rate of $\hat{\phi}_{L,M,\mathcal{H}}$ to ϕ as M increases.

Theorem 4.4.2. *Let $\mu^{\boldsymbol{x}}$ be the distribution of the initial conditions of trajectories, and $\mathcal{H}_M = \mathcal{B}_n$ with $n \asymp (M/\log M)^{\frac{1}{2s+1}}$, where $\mathcal{B}_n$ is the central ball of $\mathcal{L}_n$ with radius $c_1 + S$, and the linear space $\mathcal{L}_n \subseteq L^\infty([0,R])$ satisfies*

$$dim(\mathcal{L}_n) \leq c_0 n \quad \textit{and} \quad \inf_{\varphi \in \mathcal{L}_n} \|\varphi - \phi\|_{L^\infty} \leq c_1 n^{-s}$$

for some constants $c_0, c_1, s > 0$. Suppose that the geometric coercivity condition holds on $\mathcal{L} \coloneqq \cup_n \mathcal{L}_n$ with constant $c_{L,N,\mathcal{L},\mathcal{M}}$. Then there exists some constant $C(S,R,c_0,c_1)$ such that

$$\mathbb{E}\Big[\Big\| \hat{\phi}_{L,M,\mathcal{H}_M}(\cdot)\cdot - \phi(\cdot)\cdot \Big\|_{L^2(\rho^L_{T,\mathcal{M}})} \Big] \leq \frac{C(S,R,c_0,c_1)}{c_{L,N,\mathcal{L},\mathcal{M}}} \Big(\frac{\log M}{M} \Big)^{\frac{s}{2s+1}}.$$

The constant s is tied closely to the regularity of ϕ, and it plays an important role in the convergence rate. For example, when $\phi \in C^1$, we can take $s = 1$ with linear spaces of first degree piecewise polynomials, we end up with a $M^{\frac{1}{3}}$ learning rate. The rate is the same as the minimax rate for nonparametric regression in one dimension (up to the logarithmic factor), and is independent of the dimension $D = Nd$ of the state space. Empirical results suggest that at least in some cases, when L grows, i.e. each trajectory is sampled at more points, then the estimators improve; this is however not captured by our bound.

4.4.4 Trajectory Estimation Error

We have established the convergence of the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ to the true interaction kernel ϕ. We now establish the convergence of the trajectories of the estimated dynamics, evolved using $\hat{\phi}_{L,M,\mathcal{H}}$, to the observed trajectories.

Theorem 4.4.3. *Let $\phi \in \mathcal{K}_{R,S}$ and $\hat{\phi} \in \mathcal{K}_{R,S_0}$, for some $S_0 \geq S$. Suppose that $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$ are solutions of (4.2.1) w.r.t to ϕ and $\hat{\phi}$, respectively, for $t \in [0,T]$, with*

$\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$. *Then we have the following inequality,*

$$\mathbb{E}\Big[d_{trj}\Big(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]}\Big)^2\Big] \leq 4T^2 C(\mathcal{M}, T) \exp(64T^2 S_0^2) \left\| \phi(\cdot) \cdot - \hat{\phi}(\cdot) \cdot \right\|^2_{L^2(\rho_{T,\mathcal{M}})},$$

where $C(\mathcal{M}, T)$ is a positive constant depending only on geometric properties of $\mathcal{M}$ and T, but may be chosen independent of T if $\mathcal{M}$ is compact.

While these bounds are mainly useful for small times T, given the exponential dependence on T of the bounds, they can be overly pessimistic. It may also happen that the predicted trajectories are not accurate in terms of agent positions, but they maintain, and even predict from initial conditions, large-scale, emergent properties of the original system, such as flocking of birds of milling of fish [146]. We suspect this can hold also in the manifold setting, albeit in ways that are affected by geometric properties of the manifold.

4.5 Numerical Experiments

We consider two prototypical first order dynamics, Opinion Dynamics (OD) and Predator-Swarm dynamics (PS1), each on two different manifolds, the $2D$ sphere $\mathbb{S}^2$, centered at the origin with radius $\frac{5}{\pi}$, and the Poincaré disk $\mathbb{PD}$ (unit disk centered at the origin, with the hyperbolic metric). These are model spaces with constant positive and negative curvature, respectively. We conduct extensive experiments on these four scenarios to demonstrate the performance of the estimators both in terms of the estimation errors (approximating ϕ's) and trajectory estimator errors (estimating the observed dynamics) over $[0, T]$.

For each type of dynamics, on each of the two model manifolds, we visualize trajectories of the system, with a random initial condition (i.e. not in the training set), driven by ϕ and $\hat{\phi}$. We also augment the system by adding new agents: without any re-learning, we can transfer $\hat{\phi}$ to drive this augmented system (with $N = 40$ in

our examples), for which will also visualize the trajectories (again, started from a new random initial condition. We also report on the (relative) estimation error of the interaction kernel, as defined in (4.3.3), and on the trajectory errors, defined in (4.3.4).

For each system of $N = 20$ agents, we take $M = 500$ and $L = 500$ to generate the training data. For each $\mathcal{H}_M$, we use first-degree clamped B-splines as the basis functions with $d(\mathcal{H}_M) = \mathcal{O}(n_*) = (\frac{ML}{\log(ML)})^{\frac{1}{3}} N^{\frac{1}{d}}$. We use a geometric numerical integrator [66] (4^{th} order Backward Differentiation Formula with a projection scheme) for the evolution of the dynamics. For details, see the Appendix.

Opinion Dynamics (OD) is used to model simple interactions of opinions [6, 139] as well as choreography [26]. In fig.4.1 we display trajectories of the system on the two model manifolds. The results are summarized in fig.4.1. The relative error of the estimator $\hat{\phi}$ for OD on $\mathbb{S}^2$ is $1.894 \cdot 10^{-1} \pm 3.1 \cdot 10^{-4}$, whereas for OD on $\mathbb{PD}$ is $1.935 \cdot 10^{-1} \pm 9.5 \cdot 10^{-4}$, both are calculated using (4.3.3). The errors for trajectory prediction are reported in table 4.2.

OD	$[0, T]$
$\text{mean}_{\text{IC}}^{\mathbb{S}^2}$: Training ICs	$8.8 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$
$\text{mean}_{\text{IC}}^{\mathbb{S}^2}$: Random ICs	$9.0 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$
$\text{mean}_{\text{IC}}^{\mathbb{PD}}$: Training ICs	$1.08 \cdot 10^{-1} \pm 1.6 \cdot 10^{-3}$
$\text{mean}_{\text{IC}}^{\mathbb{PD}}$: Random ICs	$1.08 \cdot 10^{-1} \pm 2.6 \cdot 10^{-3}$

Table 4.2: (OD on $\mathbb{S}^2$ or $\mathbb{PD}$) mean_{IC} is the mean of the trajectory errors over M initial conditions (ICs), as defined in eq.(4.3.4).

Predator-Swarm System (PS1): this is a heterogeneous agent system, which is used to model interactions between multiple types of animals [32, 103]. The learning theory presented in section 4.4 is described for homogeneous agent systems, but the theory and the corresponding algorithms extend naturally to heterogeneous agent systems in a manner analogous to [90, 96]. In this case, there are different interaction kernels, $\phi_{k,k'}$, one for each (directed) interaction between agents of type k and agents of type k'. In our example here there are two types, {prey,predator}, and therefore

4 interaction kernels; however there is only one predator, so the interaction kernel predator-predator is 0. The results are visualized in fig.4.2. The (relative) errors of the estimators are in table 4.3.

$\mathrm{Err}_{1,1}^{\mathbb{S}^2} = 2.98 \cdot 10^{-1} \pm 5.9 \cdot 10^{-3}$	$\mathrm{Err}_{1,2}^{\mathbb{S}^2} = 8.4 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$
$\mathrm{Err}_{2,1}^{\mathbb{S}^2} = 2.5 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$	$\mathrm{Err}_{2,2}^{\mathbb{S}^2} = 0$
$\mathrm{Err}_{1,1}^{\mathbb{PD}} = 6.2 \cdot 10^{-2} \pm 3.0 \cdot 10^{-3}$	$\mathrm{Err}_{1,2}^{\mathbb{PD}} = 9.1 \cdot 10^{-4} \pm 4.8 \cdot 10^{-5}$
$\mathrm{Err}_{2,1}^{\mathbb{PD}} = 2.7 \cdot 10^{-3} \pm 1.4 \cdot 10^{-4}$	$\mathrm{Err}_{2,2}^{\mathbb{PD}} = 0$

Table 4.3: (PS1 on $\mathbb{S}^2$ or $\mathbb{PD}$) Relative estimation errors for $\hat{\phi}$.

PS1	$[0, T]$
$\mathrm{mean}_{\mathrm{IC}}^{\mathbb{S}^2}$: Training ICs	$2.36 \cdot 10^{-2} \pm 9.8 \cdot 10^{-4}$
$\mathrm{mean}_{\mathrm{IC}}^{\mathbb{S}^2}$: Random ICs	$2.40 \cdot 10^{-2} \pm 8.1 \cdot 10^{-4}$
$\mathrm{mean}_{\mathrm{IC}}^{\mathbb{PD}}$: Training ICs	$6.3 \cdot 10^{-3} \pm 2.0 \cdot 10^{-4}$
$\mathrm{mean}_{\mathrm{IC}}^{\mathbb{PD}}$: Random ICs	$6.4 \cdot 10^{-3} \pm 2.2 \cdot 10^{-4}$

Table 4.4: As in table 4.2, but for the PS1 system.

Discussion: As shown in the figures and tables in this section, the estimators not only provide close approximation to their corresponding interaction kernels ϕ's, but also capture additional information about the true interaction laws, e.g. the support. The accuracy on the trajectories is consistent with the theory, and the lack of overfitting and the ability to generalize well to predicting trajectories started at new random initial conditions, which in general are very far from any of the initial conditions in the training data, given the high-dimensionality of the state space. This is truly made possible because we have taken advantage of the symmetries in the system, in particular invariance of the governing equations under permutations of the agents (of the same type, in the case of heterogeneous agent systems, such as PS1), and radial symmetry of the interaction kernels. Further invariances, when the number of agents increases, make it possible to re-use the interaction kernel estimated on a system of N agents to predict trajectories of a system with the same interaction kernel, but a different number of agents, which of course has a state space of different dimension. This admittedly simple example of transfer would not possible for general-purpose techniques that directly estimate the r.h.s. of the system of ODEs.

4.6 Conclusion

We have considered the problem of estimating the dynamics of a special yet widely used set of dynamical systems, consisting of interacting agents on Riemannian manifolds. These are driven by a first-order system of ODEs on the manifold, with a typically very high-dimensional state space $\mathcal{M}^N$, where N is the (typically large) number of agents. We constructed estimators that are optimal and avoid the curse of dimensionality, but exploiting the multiple symmetries in these systems, and the simplicity of the underlying interaction laws. Extensions to more complex systems of interacting agents may be considered, in particular to second-order systems, which will require the use of parallel transport on $\mathcal{M}$, to more general interaction kernels, depending on other variables beyond pairwise distances, as well as to systems interacting with a varying environment.

4.7 Preliminaries

In this work, $\mathcal{M}$ is a connected, smooth, and geodesically complete d-dimensional Riemannian manifold with Riemannian metric g. For details regarding the basic definitions of Riemannian manifolds, geodesics, Riemannian distances, exponential maps, cut loci, and injectivity radii, please see [79, 56]. We will discuss how to find the minimal geodesic and the Riemannian distance between any two points on the two prototypical manifolds used in our numerical algorithms: the two-dimensional sphere ($\mathbb{S}^2$) and the Poincaré Disk ($\mathbb{PD}$).

4.7.1 Riemannian Geometry on the $2D$ Sphere

The $2D$ Sphere ($\mathbb{S}^2$) of radius r and centered at the origin can be isometrically embedded in $\mathbb{R}^3$ in the natural way, i.e., $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^2 \subset \mathbb{R}^3$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^2$, the

Riemannian distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$d_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{y}) = r \cdot \theta, \quad \theta = \operatorname{acos}\left(\frac{\langle \boldsymbol{x},\boldsymbol{y},\rangle}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|}\right).$$

The minimal geodesic between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the piece of the arc on the great circle of $\mathbb{S}^2$ with the smallest length, assuming $\boldsymbol{x}$ and $\boldsymbol{y}$ are not in each others' cut locus, i.e. diametrically opposed. The unit vector on the minimal geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$, denoted as $\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y})$, can be computed as follows

$$\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{y}-\boldsymbol{x}-\operatorname{Proj}_{-\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{x})}{\left\|\boldsymbol{y}-\boldsymbol{x}-\operatorname{Proj}_{-\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{x})\right\|}.$$

Here $\operatorname{Proj}_{\boldsymbol{u}}(\boldsymbol{w})$ is the projection of $\boldsymbol{w}$ onto $\boldsymbol{u}$.

4.7.2 Riemannian Geometry on the Poincaré Disk

For any two points $\boldsymbol{x},\boldsymbol{y} \in \mathbb{PD}$ on the Poincaré Disk ($\mathbb{PD}$) where $\mathbb{PD} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^2 \text{ s.t. } \|\boldsymbol{x}\| < 1\}$, the Riemannian metric, written in the standard coordinates of $\mathbb{R}^2$, is given by

$$g_{i,j}(\boldsymbol{x}) = \frac{4\delta_{i,j}}{(1-\|\boldsymbol{x}\|^2)^2}, \quad \boldsymbol{x} \in \mathbb{PD},$$

with $\delta_{i,j}$ being the Kronecker delta, and the corresponding Riemannian distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$d_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{y}) = \operatorname{acosh}\left(1 + \frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{(1-\|\boldsymbol{x}\|^2)(1-\|\boldsymbol{y}\|^2)}\right).$$

The minimal geodesics between $\boldsymbol{x}$ and $\boldsymbol{y}$ are either straight line segments if $\boldsymbol{x}$ and $\boldsymbol{y}$ are on a line through the origin or circular arc perpendicular to the boundary. For the straight line segment case, we have the unit vector on the minimal geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$, denoted as $\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y})$, computed as follows: we identify the vector $\boldsymbol{y}-\boldsymbol{x}$, computed

in $\mathbb{R}^2$ as a tangent vector in $T_{\boldsymbol{x}}\mathcal{M}$, then normalize it to obtain $\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{y}-\boldsymbol{x}}{\|\boldsymbol{y}-\boldsymbol{x}\|_{T_{\boldsymbol{x}}\mathcal{M}}}$. For the perpendicular arc case, we first find the inverse $\boldsymbol{y}'$ of $\boldsymbol{y}$ w.r.t to the unit disk (in $\mathbb{R}^2$); then we use the three points $\boldsymbol{x},\boldsymbol{y},\boldsymbol{y}'$ to find the center $\boldsymbol{o}'$ of the circle passing through $\boldsymbol{x},\boldsymbol{y}$ and $\boldsymbol{y}'$. Then the unit tangent vector on the geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$ is computed as follows: , we compute $\boldsymbol{y}-\boldsymbol{x}-\mathrm{Proj}_{\boldsymbol{o}'-\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{x})$ in $\mathbb{R}^2$ (with the Euclidean metric), then identify it as a tangent vector in $T_{\boldsymbol{x}}\mathcal{M}$, and normalize it:

$$\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{y}-\boldsymbol{x}-\mathrm{Proj}_{\boldsymbol{o}'-\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{x})}{\left\|\boldsymbol{y}-\boldsymbol{x}-\mathrm{Proj}_{\boldsymbol{o}'-\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{x})\right\|_{T_{\boldsymbol{x}}\mathcal{M}}}.$$

4.8 Learning Theory: Foundation

In this section, we present the theoretical foundation needed to prove the theorems presented in the main body. We follow the ideas presented in [89] with similar strategies presented in [44, 64]. We begin with the following assumption.

Assumption 4.8.1. *$\mathcal{H}$ is a compact (in L^∞-norm) and convex subset of $L^2([0,R])$, such that every $\varphi\in\mathcal{H}$ is bounded above by some constant $S_0\geq S$, i.e. $\|\varphi\|_{L^\infty([0,R])}\leq S_0$; moreover φ is smooth enough to ensure the existence and uniqueness of solutions of*

$$\dot{\boldsymbol{x}}_i(t) = \frac{1}{N}\sum_{i'=1}^{N}\phi(d_{\mathcal{M}}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t)))\boldsymbol{w}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t)),\qquad i=1,\dots,N. \tag{4.8.1}$$

for $t\in[0,T]$, i.e. $\varphi\in\mathcal{H}\cap\mathcal{K}_{R,S_0}$.

Another important observation is that since $\phi\in\mathcal{K}_{R,S}$ and T is finite, the distribution of $\boldsymbol{x}_i(t)$'s does not blow up over $[0,T]$ ensuring that the $\boldsymbol{x}_i(t)$'s have bounded distance from the $\boldsymbol{x}_i(0)$'s. In fact, let R_0 be the maximum Riemannian distance

between any pair of agents at $t = 0$, then

$$\max_{i,i'=1,\dots,N} r_{i,i'}(t) = \max_{i,i'=1,\dots,N} d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)) \leq R_0 + TRS, \quad \text{for } t \in [0,T].$$

Hence the $\boldsymbol{x}_i(t)$'s live in a compact (w.r.t to the $d_{\mathcal{M}}$ metric) ball around the $\boldsymbol{x}_i(0)$'s, denoted as $\mathcal{B}_{\mathcal{M}}(\boldsymbol{X}_0, R_1)$ where $R_1 = R_0 + TRS$. Recall the definition of the loss functional used to find the estimator, namely $\hat{\phi}_{L,M,\mathcal{H}}$ to the unknown interaction kernel ϕ, give by

$$\mathcal{E}_{L,M,\mathcal{M}}(\varphi) := \frac{1}{ML} \sum_{l,m=1}^{L,M} \left\| \dot{\boldsymbol{X}}_{t_l}^m - \boldsymbol{f}_{\varphi}^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m) \right\|_{T_{\boldsymbol{X}_{t_l}^m}\mathcal{M}^N}^2 . \tag{4.8.2}$$

Further recall that the estimator is defined as $\hat{\phi}_{L,M,\mathcal{H}} := \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,M,\mathcal{M}}(\varphi)$. When $M \to \infty$, we obtain the following loss functional (by the law of large numbers).

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) := \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{X}_0 \sim \mu^{\boldsymbol{x}}} \Big[\left\| \dot{\boldsymbol{X}}_{t_l} - \boldsymbol{f}_{\varphi}^{\mathrm{c}}(\boldsymbol{X}_{t_l}) \right\|_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}^2 \Big]. \tag{4.8.3}$$

The minimizer of $\mathcal{E}_{L,\infty,\mathcal{M}}$ over $\mathcal{H}$ is defined as $\hat{\phi}_{L,\infty,\mathcal{H}}$, which is closely related to $\hat{\phi}_{L,M,\mathcal{H}}$ (in the $M \to \infty$ sense). And they are close to ϕ, when we establish the following condition on $\mathcal{H}$.

Definition 4.8.1 (Geometric Coercivity condition)**.** *The geometric evolution system in* (4.8.1) *with initial condition sampled from $\mu^{\boldsymbol{x}}$ on $\mathcal{M}^N$ is said to satisfy the geometric coercivity condition on the admissible hypothesis space $\mathcal{H}$ if there exists a constant $c_{L,N,\mathcal{H},\mathcal{M}} > 0$ such that for any $\varphi \in \mathcal{H}$ with $\varphi(\cdot)\cdot \in L^2(\rho_{T,\mathcal{M}}^L)$, the following inequality holds:*

$$c_{L,N,\mathcal{H},\mathcal{M}} \left\| \varphi(\cdot)\cdot \right\|_{L^2(\rho_{T,\mathcal{M}}^L)}^2 \leq \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{X}_0 \sim \mu^{\boldsymbol{x}}} \Big[\left\| \boldsymbol{f}_{\varphi}^{c}(\boldsymbol{X}_{t_l}) \right\|_{T_{\boldsymbol{X}_{t_l}}\mathcal{M}^N}^2 \Big]. \tag{4.8.4}$$

From this condition, we can derive the following theorem.

Theorem 4.8.2. *Let $\phi \in L^2([0,R])$, and $\mathcal{H}$ a compact (w.r.t the L^∞ norm) and convex subset of $L^2([0,R])$ such that the geometric coercivity condition* (4.8.4) *holds with a constant $c_{L,N,\mathcal{H},\mathcal{M}}$. Then, for $\hat{\phi}_{L,M,\mathcal{H}}$, estimated by minimizing* (4.8.2) *on the trajectory data generated by* (4.8.1)*, the following inequality*

$$\left\|\hat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\right\|^2_{L^2(\rho^L_{T,\mathcal{M}})} \leq \frac{2}{c_{L,N,\mathcal{H},\mathcal{M}}}\Big(\epsilon + \inf_{\varphi\in\mathcal{H}} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho^L_{T,\mathcal{M}})}\Big) \tag{4.8.5}$$

holds with probability at least $1-\tau$, when $M \geq \frac{1152 S_0^2 R^2}{\epsilon c_{L,N,\mathcal{H},\mathcal{M}}}\Big(\ln(\mathcal{N}(\mathcal{H}, \frac{\epsilon}{48 S_0 R^2})) + \ln(\frac{1}{\tau})\Big)$. Here $\mathcal{N}(\mathcal{U},\epsilon)$ is the covering number of a set $\mathcal{U}$ with open balls of radius ϵ w.r.t the L^∞-norm.

Using this concentration result, we can get the strong consistency of our estimators under mild hypotheses.

Theorem 4.8.3. *For a family of compact (w.r.t. the L^∞ norm) convex subsets, $\{\mathcal{H}_M\}_{M=1}^\infty$, of $L^2([0,R])$, when the following conditions hold, (i) $\cup_M \mathcal{H}_M$ is compact in L^∞; (ii) the geometric coercivity condition,* (4.8.1)*, holds on $\cup_M \mathcal{H}_M$; (iii) $\inf_{\varphi\in\mathcal{H}_M} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})} \overset{M\to\infty}{\longrightarrow} 0$, then*

$$\lim_{M\to\infty} \left\|\hat{\phi}_{L,M,\mathcal{H}_M}(\cdot)\cdot - \phi(\cdot)\cdot\right\|_{L^2(\rho^L_{T,\mathcal{M}})} = 0 \quad a.s. \tag{4.8.6}$$

This theorem establishes the almost sure convergence of our estimator to the true interaction kernel as $M\to\infty$.

4.8.1 Concentration and Consistency

Our first step is to establish the consistency of the estimator for the true kernel ϕ of the system. Note that $\mathcal{H}$ can be embedded as a compact (in L^∞ sense) set of $L^2(\rho^L_{T,\mathcal{M}})$.

We establish a strong consistency result on our estimators of the form,

$$\lim_{M\to\infty} \left\| \hat{\phi}_{L,M}(\cdot)\cdot - \phi(\cdot)\cdot \right\|_{L^2(\rho^L_{T,\mathcal{M}})} = 0, \ a.s.$$

Our discussions of consistency under the L^2-norm on manifolds can be regarded as a natural extension from the case on Euclidean Space in [89]. We define the following loss functional of the vectorized system, $\boldsymbol{X}_t$

$$\begin{aligned}\mathcal{E}_{\boldsymbol{X}_t}(\varphi) &:= \frac{1}{N}\sum_{i=1}^N \left\| \frac{1}{N}\sum_{i'=1}^N (\phi_{ii',t} - \varphi_{ii',t})\boldsymbol{w}_{ii',t} \right\|^2_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \\ &= \frac{1}{N}\sum_{i=1}^N \langle \frac{1}{N}\sum_{i'=1}^N (\phi_{ii',t} - \varphi_{ii',t})\boldsymbol{w}_{ii',t}, \frac{1}{N}\sum_{i''=1}^N (\phi_{ii'',t} - \varphi_{ii'',t})\boldsymbol{w}_{ii'',t}, \rangle \, g(\boldsymbol{x}_i(t)).\end{aligned} \tag{4.8.7}$$

Here we take $\boldsymbol{w}_{ii',t} = d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))\boldsymbol{v}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ and $\phi_{ii',t} = \phi(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$; similarly for $\varphi_{ii',t}$. Now we can see that

$$\mathcal{E}_{L,M,\mathcal{M}}(\varphi) = \frac{1}{LM}\sum_{l,m=1}^{L,M} \mathcal{E}_{\boldsymbol{X}^m_{t_l}}(\varphi).$$

When $M \to \infty$, we have, by the law of large numbers,

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) = \frac{1}{L}\sum_{l=1}^L \mathbb{E}_{\boldsymbol{X}_0\sim\mu^{\boldsymbol{x}}} \mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi)\,.$$

We are ready to summarize some basic properties of $\mathcal{E}_{\boldsymbol{X}_t}(\varphi)$.

Proposition 4.1. *For $\varphi_1, \varphi_2 \in \mathcal{H}$, we have*

$$\left|\mathcal{E}_{\boldsymbol{X}_t}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_t}(\varphi_2)\right| \le \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})} \, \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}\,. \tag{4.8.8}$$

Here we define the probability measure, $\hat{\rho}^t_{\mathcal{M}}(r) \coloneqq \frac{1}{N^2}\sum_{i,i'=1}^N \delta_{d_{\mathcal{M}}(\boldsymbol{x}_i(t),\boldsymbol{x}_{i'}(t))}(r)$.

Proof. Let $\varphi_1, \varphi_2 \in \mathcal{H}$, and define $\varphi^1_{ii',t} \coloneqq \varphi_1(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)))$, similarly for $\varphi^2_{ii',t}$. Moreover, let $r_{ii',t} \coloneqq d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ and $\boldsymbol{w}_{ii',t} \coloneqq d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))\boldsymbol{v}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$. Immediately, we have

$$\|\boldsymbol{w}_{ii',t}\|_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \leq r_{ii',t},$$

since $\boldsymbol{v}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ has either length 1 or 0. Next, using Jensen's inequality, we have

$$\begin{aligned}
\left|\mathcal{E}_{\boldsymbol{X}_t}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_t}(\varphi_2)\right| = \Bigg| \frac{1}{N}\sum_{i=1}^{N} \langle \frac{1}{N}\sum_{i'=1}^{N} (\varphi^1_{ii',t} - \varphi^2_{ii',t})\boldsymbol{w}_{ii',t}, \\
\frac{1}{N}\sum_{i''=1}^{N} (2\phi_{ii'',t} - \varphi^1_{ii'',t} - \varphi^2_{ii'',t})\boldsymbol{w}_{ii'',t}, \rangle \, g(\boldsymbol{x}_i(t)) \Bigg| \\
\leq \frac{1}{N}\sum_{i=1}^{N} \left\| \frac{1}{N}\sum_{i'=1}^{N} (\varphi^1_{ii',t} - \varphi^2_{ii',t})\boldsymbol{w}_{ii',t} \right\|_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \\
\left\| \frac{1}{N}\sum_{i''=1}^{N} (2\phi_{ii'',t} - \varphi^1_{ii'',t} - \varphi^2_{ii'',t})\boldsymbol{w}_{ii'',t} \right\|_{T_{\boldsymbol{x}_i(t)}\mathcal{M}} \\
\leq \sqrt{\frac{1}{N^2}\sum_{i,i'=1}^{N} (\varphi^1_{ii',t} - \varphi^2_{ii',t}) r^2_{ii',t}} \sqrt{\frac{1}{N^2}\sum_{i,i''=1}^{N} (2\phi_{ii'',t} - \varphi^1_{ii',t} - \varphi^2_{ii',t}) r^2_{ii'',t}} \\
\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot)\|_{L^2(\hat{\rho}^t_{\mathcal{M}})} \, \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})},
\end{aligned}$$

where $\hat{\rho}^t_{\mathcal{M}}(r) = \frac{1}{N^2}\sum_{i,i'=1}^{N} \delta_{r_{ii',t}}(r)$. □

With Proposition 4.1 proven, we get the following proposition establishing the continuity of our error functionals.

Proposition 4.2. *For $\varphi_1, \varphi_2 \in \mathcal{H}$, we have the inequalities*

$$\begin{aligned}
\left|\mathcal{E}_{L,M,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi_2)\right| &\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty} \, \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty} \\
\left|\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_2)\right| &\leq \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})} \, \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})}.
\end{aligned} \tag{4.8.9}$$

Proof. Using the results from Prop. 4.1, and defining $\hat{\rho}^L_{T,\mathcal{M}} \coloneqq \frac{1}{L}\sum_{l=1}^L \hat{\rho}^{t_l}_{\mathcal{M}}$, we have

$$\begin{aligned}
&\big|\frac{1}{L}\sum_{l=1}^L \mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_1) - \frac{1}{L}\sum_{l=1}^L \mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_2)\big| \le \frac{1}{L}\sum_{l=1}^L \big|\mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_1) - \mathcal{E}_{\boldsymbol{X}_{t_l}}(\varphi_2)\big| \\
&< \frac{1}{L}\sum_{l=1}^L \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})} \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})} \\
&\le \sqrt{\frac{1}{L}\sum_{1=1}^L \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}} \sqrt{\frac{1}{L}\sum_{l=1}^L \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^t_{\mathcal{M}})}} \\
&= \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})} \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})}
\end{aligned}$$

Next, we have

$$\begin{aligned}
\big|\mathcal{E}_{L,M,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi_2)\big| &\le \frac{1}{M}\sum_{m=1}^M \big|\frac{1}{L}\sum_{l=1}^L \mathcal{E}_{\boldsymbol{X}^m_{t_l}}(\varphi_1) - \frac{1}{L}\sum_{l=1}^L \mathcal{E}_{\boldsymbol{X}^m_{t_l}}(\varphi_2)\big| \\
&\le \frac{1}{M}\sum_{m=1}^M \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})} \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\hat{\rho}^L_{T,\mathcal{M}})} \\
&\le \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty} \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^\infty} \\
&\le R^2 \|\varphi_1 - \varphi_2\|_{L^\infty} \|2\phi - \varphi_1 - \varphi_2\|_{L^\infty} \ .
\end{aligned}$$

Meanwhile, taking $M \to \infty$ for $\big|\mathcal{E}_{L,M,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi_2)\big|$, we obtain

$$\big|\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_1) - \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi_2)\big| \le \|\varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})} \|2\phi(\cdot)\cdot - \varphi_1(\cdot)\cdot - \varphi_2(\cdot)\cdot\|_{L^2(\rho^L_{T,\mathcal{M}})} ,$$

where $\rho^L_{T,\mathcal{M}} = \mathbb{E}_{\boldsymbol{X}_0\sim\mu^{\boldsymbol{x}}}[\hat{\rho}^L_{T,\mathcal{M}}]$. □

As a further derivation, we observe that for any $\varphi \in \mathcal{H} \subset L^2([0,R])$, we have that $\max_{r\in[0,R]} \big|\varphi(\cdot)\cdot\big| \le R \max_{r\in[0,R]} \big|\varphi(\cdot)\big|$, so we obtain the following Corollary:

Corollary 4.8.4. *For $\varphi \in \mathcal{H}$, define*

$$\mathcal{L}_M(\psi) \coloneqq \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) - \mathcal{E}_{L,M,\mathcal{M}}(\varphi),$$

then for any $\varphi_1, \varphi_2 \in \mathcal{H}$, we have

$$\left|\mathcal{L}_M(\varphi_1) - \mathcal{L}_M(\varphi_2)\right| \leq 2R^2 \left\|\varphi_1 - \varphi_2\right\|_{L^\infty} \left\|2\phi - \varphi_1 - \varphi_2\right\|_{L^\infty}.$$

Now we can consider the distance between the minimizer of the error functional $\mathcal{E}_{L,\infty,\mathcal{M}}$ over $\mathcal{H}$ and any other $\varphi \in \mathcal{H}$. Let

$$\hat{\phi}_{L,\infty,\mathcal{H}} = \arg\min_{\varphi \in \mathcal{H}} \mathcal{E}_{L,\infty,\mathcal{M}}(\varphi).$$

From the geometric coercivity condition and the convexity of $\mathcal{H}$, we obtain

Proposition 4.3. *For any $\varphi \in \mathcal{H}$,*

$$\mathcal{E}_{L,\infty,\mathcal{M}}(\varphi) - \mathcal{E}_{L,\infty,\mathcal{M}}(\hat{\phi}_{L,\infty,\mathcal{H}}) \geq c_{L,N,\mathcal{H},\mathcal{M}} \left\|\varphi(\cdot)\cdot - \hat{\phi}_{L,\infty,\mathcal{H}}(\cdot)\cdot\right\|_{L^2(\rho^L_{T,\mathcal{M}})}. \tag{4.8.10}$$

We now define the defect function $\mathcal{D}_{L,M,\mathcal{H}}(\varphi) \coloneqq \mathcal{E}_{L,M,\mathcal{M}}(\varphi) - \mathcal{E}_{L,M,\mathcal{M}}(\hat{\phi}_{L,\infty,\mathcal{H}})$, and define

$$\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) \coloneqq \lim_{M\to\infty} \mathcal{D}_{L,M,\mathcal{H}}(\varphi) = \mathcal{E}_{L,\infty,\mathcal{H}}(\varphi) - \mathcal{E}_{L,\infty,\mathcal{M}}(\hat{\phi}_{L,\infty,\mathcal{H}}).$$

Then, we show that we can uniformly bound $\frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\cdot) - \mathcal{D}_{L,M,\mathcal{H}}(\cdot)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\cdot)+\epsilon}$ on $\mathcal{H}$ with high probability,

Proposition 4.4. *For any $\epsilon > 0$ and $\alpha \in (0,1)$, we have*

$$\mathbb{P}_{\mu^{\boldsymbol{x}}}\left(\sup_{\varphi\in\mathcal{H}} \frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi)+\epsilon} \geq 3\alpha\right) \leq \mathcal{N}\Big(\mathcal{H}, \frac{\alpha\epsilon}{8S_0R^2}\Big) \exp\left(-\frac{c_{L,N,\mathcal{H},\mathcal{M}}\alpha^2 M\epsilon}{32S_0^2}\right)$$

where $\mathcal{N}(U,r)$ is the covering number of set U with open balls of radius r w.r.t the $L^\infty-$norm.

The proof of Proposition 4.4 uses the following Lemma similar to Lemma 19 in [89],

Lemma 4.8.5. *For any $\epsilon > 0$ and $\alpha \in (0,1)$, if $\varphi_1 \in \mathcal{H}$ satisfies*

$$\frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_1) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi_1)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_1) + \epsilon} < \alpha$$

then for any $\varphi_2 \in \mathcal{H}$ s.t. $\|\varphi_1 - \varphi_2\|_{L^\infty} \leq r_0 = \frac{\alpha\epsilon}{8S_0R^2}$, we have

$$\frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_2) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi_2)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi_2) + \epsilon} < 3\alpha$$

Using the results we have just established, the proofs of theorems 4.8.2 and 4.8.3 now follow similarly to the analogous results in [89, 90, 96].

4.8.2 Rate of Convergence

Using these results, we establish the convergence rate of $\hat{\phi}_{L,M,\mathcal{H}}$ to ϕ as M increases.

Theorem 4.8.6. *Let $\mu^{\boldsymbol{x}}$ be the distribution of the initial conditions of trajectories, and $\mathcal{H}_M = \mathcal{B}_n$ with $n \asymp (M/\log M)^{\frac{1}{2s+1}}$, where $\mathcal{B}_n$ is the central ball of $\mathcal{L}_n$ with radius $c_1 + S$, and the linear space $\mathcal{L}_n \subseteq L^\infty([0,R])$ satisfies the dimension and approximation conditions below,*

$$dim(\mathcal{L}_n) \leq c_0 n \quad \text{and} \quad \inf_{\varphi \in \mathcal{L}_n} \|\varphi - \phi\|_{L^\infty} \leq c_1 n^{-s}$$

for some constants $c_0, c_1, s > 0$. Suppose that the geometric coercivity condition holds on $\mathcal{L} \coloneqq \cup_n \mathcal{L}_n$ with constant $c_{L,N,\mathcal{L},\mathcal{M}}$. Then there exists some constant $C(S,R,c_0,c_1)$ such that

$$\mathbb{E}\Big[\Big\| \hat{\phi}_{L,M,\mathcal{H}_M}(\cdot)\cdot - \phi(\cdot)\cdot \Big\|_{L^2(\rho^L_{T,\mathcal{M}})} \Big] \leq \frac{C(S,R,c_0,c_1)}{c_{L,N,\mathcal{L},\mathcal{M}}} \Big(\frac{\log M}{M} \Big)^{\frac{s}{2s+1}}.$$

The proof of the theorem closely follows the ideas in [89] and their further development in [90, 96], and is therefore omitted.

4.8.3 Trajectory Estimation Error

Recall the following theorem on the trajectory estimator error:

Theorem 4.8.7. *Let $\phi \in \mathcal{K}_{R,S}$ and $\hat{\phi} \in \mathcal{K}_{R,S_0}$, for some $S_0 \geq S$. Suppose that $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$ are solutions of (4.8.1) w.r.t to ϕ and $\hat{\phi}$, respectively, for $t \in [0,T]$, with $\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$. Then the following inequalities hold:*

$$d_{traj,\mathcal{M}^N}\left(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]}\right)^2 \leq 4TC(\mathcal{M},T)\exp(64T^2S_0^2)\left\|\dot{\boldsymbol{X}}_t - \boldsymbol{f}_{\hat{\phi}}^c(\boldsymbol{X}_t)\right\|^2_{T_{\boldsymbol{X}_t}M^N}, \tag{4.8.11}$$

and

$$\mathbb{E}_{\boldsymbol{X}_0\sim\mu^{\boldsymbol{x}}}\left[d_{traj,\mathcal{M}^N}\left(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]}\right)^2\right] \leq 4T^2C(\mathcal{M},T)\exp(64T^2S_0^2)\left\|\phi(\cdot)\cdot - \hat{\phi}(\cdot)\cdot\right\|^2_{L^2(\rho_{T,\mathcal{M}})}, \tag{4.8.12}$$

where $C(\mathcal{M},T)$ is a positive constant depending only on geometric properties of $\mathcal{M}$ and T, but may be chosen independent of T if $\mathcal{M}$ is compact.

It states two different estimates of the trajectory estimation error. First, it bounds the system trajectory error for any one single initial condition; second, it bounds the expectation of the worst trajectory estimation error on time interval $[0,T]$ among all different initial conditions.

Proof of Theorem 4.8.7. Assume that $\phi \in \mathcal{K}_{R,S}$, $\hat{\phi} \in \mathcal{K}_{R,S_0}$, and $\boldsymbol{X}_t, \hat{\boldsymbol{X}}_t$ are two system states generated by $\phi, \hat{\phi}$ with the same initial conditions at some $t \in [0,T]$. Next, we assume that $\mathcal{M}$ is isometrically embedded in $\mathbb{R}^{d'}$ (at least one such embedding exists, by Nash's embedding theorem), via a map $\mathcal{I}: \mathcal{M} \to \mathbb{R}^{d'}$. From now on, we will identify $\boldsymbol{x}_i$ with $\mathcal{I}\boldsymbol{x}_i$. Then for any $t \in [0,T]$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t)\|^2_{\mathbb{R}^{d'}} &= \frac{1}{N}\sum_{i=1}^N \left\|\int_{s=0}^t (\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s))ds\right\|^2_{\mathbb{R}^{d'}} \\ &\leq \frac{1}{N}\sum_{i=1}^N t\int_{s=0}^t \|\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s)\|^2_{\mathbb{R}^{d'}} ds\end{aligned}$$

$$\leq \frac{T}{N} \sum_{i=1}^{N} \int_{s=0}^{t} \|\dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s)\|_{\mathbb{R}^{d'}}^2 ds.$$

Define the function $F_{\varphi}^{\mathcal{M}}(\boldsymbol{x}, \cdot) : \mathcal{M} \to T_{\boldsymbol{x}}\mathcal{M}$ for every $\boldsymbol{x} \in \mathcal{M}$ as $F_{\varphi}^{\mathcal{M}}(\boldsymbol{x}, \cdot) \coloneqq \varphi(d_{\mathcal{M}}(\boldsymbol{x}, \cdot))\boldsymbol{w}(\boldsymbol{x}, \cdot)$. Let $F_{\varphi,ii',t}^{\mathcal{M}} = F_{\varphi}^{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t))$ and $F_{\varphi,\hat{i}\hat{i}',t}^{\mathcal{M}} = F_{\varphi}^{\mathcal{M}}(\hat{\boldsymbol{x}}_i(t), \hat{\boldsymbol{x}}_{i'}(t))$. Then

$$\begin{aligned}
&\sum_{i=1}^{N} \int_{s=0}^{t} \left\| \dot{\boldsymbol{x}}_i(s) - \dot{\hat{\boldsymbol{x}}}_i(s) \right\|_{\mathbb{R}^{d'}}^2 ds = \sum_{i=1}^{N} \int_{s=0}^{t} \left\| \dot{\boldsymbol{x}}_i(s) - \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},\hat{i}\hat{i}',s}^{\mathcal{M}} \right\|_{\mathbb{R}^{d'}}^2 ds \\
&\leq 2 \sum_{i=1}^{N} \int_{s=0}^{t} \Big(\left\| \dot{\boldsymbol{x}}_i(s) - \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},ii',s}^{\mathcal{M}} \right\|_{\mathbb{R}^{d'}}^2 + \left\| \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},ii',s}^{\mathcal{M}} - \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},\hat{i}\hat{i}',s}^{\mathcal{M}} \right\|_{\mathbb{R}^{d'}}^2 \Big) ds \\
&= 2 \sum_{i=1}^{N} \int_{s=0}^{t} \Big(\left\| \dot{\boldsymbol{x}}_i(s) - \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},ii',s}^{\mathcal{M}} \right\|_{\mathbb{R}^{d'}}^2 + I(s) \Big) ds \,.
\end{aligned}$$

Next,

$$\begin{aligned}
I(s) &= \left\| \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},ii',s}^{\mathcal{M}} - \frac{1}{N} \sum_{i'=1}^{N} F_{\hat{\phi},\hat{i}\hat{i}',s}^{\mathcal{M}} \right\|_{\mathbb{R}^{d'}}^2 = \frac{1}{N^2} \left\| \sum_{i'=1}^{N} \big(F_{\hat{\phi},ii',s}^{\mathcal{M}} - F_{\hat{\phi},ii',s}^{\mathcal{M}} + F_{\hat{\phi},ii',s}^{\mathcal{M}} - F_{\hat{\phi},\hat{i}\hat{i}',s}^{\mathcal{M}} \big) \right\|_{\mathbb{R}^{d'}}^2 \\
&\leq \frac{2}{N^2} \Big(\left\| \sum_{i'=1}^{N} \big(F_{\hat{\phi},ii',s}^{\mathcal{M}} - F_{\hat{\phi},i\hat{i}',s}^{\mathcal{M}} \big) \right\|_{\mathbb{R}^{d'}}^2 + \left\| \sum_{i'=1}^{N} \big(F_{\hat{\phi},i\hat{i}',s}^{\mathcal{M}} - F_{\hat{\phi},\hat{i}\hat{i}',s}^{\mathcal{M}} \big) \right\|_{\mathbb{R}^{d'}}^2 \Big) .
\end{aligned}$$

Since $\hat{\phi} \in \mathcal{K}_{R,S_0}$, $F_{\hat{\phi}}^{\mathcal{M}}$ is Lipschitz in each of its arguments; moreover, $\max_{r\in[0,R]} \left|\hat{\phi}\right| \leq S_0$, so that $\mathrm{Lip}(F_{\hat{\phi}}^{\mathcal{M}}(\boldsymbol{x}, \cdot)), \mathrm{Lip}(F_{\hat{\phi}}^{\mathcal{M}}(\cdot, \boldsymbol{x})) \leq 2S_0$. Therefore,

$$\begin{aligned}
I(s) &\leq \frac{2}{N^2} \Big(2\mathrm{Lip}(F_{\hat{\phi}}^{\mathcal{M}}(\boldsymbol{x}_i(s), \cdot))^2 \sum_{i'=1}^{N} \|\boldsymbol{x}_{i'}(s) - \hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2 \\
&\quad + 2 \sum_{i'=1}^{N} \mathrm{Lip}(F_{\hat{\phi}}^{\mathcal{M}}(\cdot, \hat{\boldsymbol{x}}_{i'}(s)))^2 \|\boldsymbol{x}_i(s) - \hat{\boldsymbol{x}}_i(s)\|_{\mathbb{R}^{d'}}^2 \Big) \\
&\leq \frac{4}{N^2} \mathrm{Lip}(F_{\hat{\phi}}^{\mathcal{M}}(\boldsymbol{x}_i(s), \cdot))^2 \sum_{i'=1}^{N} \|\boldsymbol{x}_{i'}(s) - \hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2 \\
&\quad + \frac{4}{N^2} \sum_{i'=1}^{N} \mathrm{Lip}(F_{\hat{\phi}}^{\mathcal{M}}(\cdot, \hat{\boldsymbol{x}}_{i'}(s)))^2 \|\boldsymbol{x}_{i'}(s) - \hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2
\end{aligned}$$

$$\leq \frac{16S_0^2}{N^2}\sum_{i'=1}^{N}\|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2+\frac{16S_0^2}{N^2}\sum_{i'=1}^{N}\|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2$$
$$\leq \frac{32S_0^2}{N^2}\sum_{i'=1}^{N}\|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2 \ .$$

Putting these results together, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|_{\mathbb{R}^{d'}}^2 &\leq \frac{2T}{N}\sum_{i=1}^{N}\int_{s=0}^{t}\Big(\Big\|\dot{\boldsymbol{x}}_i(s)-\frac{1}{N}\sum_{i'=1}^{N}F^{\mathcal{M}}_{\hat{\phi},ii',s}\Big\|_{\mathbb{R}^{d'}}^2\\ &\quad+\frac{32S_0^2}{N^2}\sum_{i'=1}^{N}\|\boldsymbol{x}_{i'}(s)-\hat{\boldsymbol{x}}_{i'}(s)\|_{\mathbb{R}^{d'}}^2\Big)\,ds\\ &=\frac{64TS_0^2}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|_{\mathbb{R}^{d'}}^2\\ &\quad+\frac{2T}{N}\sum_{i=1}^{N}\int_{s=0}^{t}\Big\|\dot{\boldsymbol{x}}_i(s)-\frac{1}{N}\sum_{i'=1}^{N}F^{\mathcal{M}}_{\hat{\phi},ii',s}\Big\|_{\mathbb{R}^{d'}}^2 ds.\end{aligned}$$

By Grönwall's inequality, we have

$$\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|_{\mathbb{R}^{d'}}^2 \leq \frac{2T}{N}\exp(64T^2S_0^2)\sum_{i=1}^{N}\int_{s=0}^{t}\Big\|\dot{\boldsymbol{x}}_i(s)-\frac{1}{N}\sum_{i'=1}^{N}F^{\mathcal{M}}_{\hat{\phi},ii',s}\Big\|_{\mathbb{R}^{d'}}^2 ds.$$

Recall that T is small, hence the solution $\boldsymbol{X}_t$ and $\hat{\boldsymbol{X}}_t$ live in a compact neighborhood of the initial condition, $\boldsymbol{X}_0=\hat{\boldsymbol{X}}_0\in\mathcal{M}^N$; i.e. $\boldsymbol{X}_t,\hat{\boldsymbol{X}}_t\in\mathcal{B}_{\mathcal{M}}(\boldsymbol{X}_0,R_2)$ with $R_2=R_0+TRS_0$. From the compactness of (the closure of) this set, and via the embedding $\mathcal{I}$, we deduce that there exists a constant $C_1(\mathcal{M},\mathcal{I},T)$ such that

$$d_{\mathcal{M}}(\boldsymbol{x}_i(t),\hat{\boldsymbol{x}}_i(t))\leq C_1(\mathcal{M},\mathcal{I},T)\,\|\boldsymbol{x}_i(t)-\hat{\boldsymbol{x}}_i(t)\|_{\mathbb{R}^{d'}}\,,\quad \text{for } t\in[0,T].$$

Since $\mathcal{I}$ is isometric, for $\boldsymbol{u}\in T_{\boldsymbol{x}}\mathcal{M}$ we have $\|d\mathcal{I}(\boldsymbol{u})\|_{\mathbb{R}^{d'}}=\|\boldsymbol{u}\|_{T_{\boldsymbol{x}}\mathcal{M}}$. Using both the

bounds above, we have

$$\begin{aligned} d_{\mathcal{M}}(\boldsymbol{X}_t, \hat{\boldsymbol{X}}_t)^2 &= \frac{1}{N}\sum_{i=1}^N d_{\mathcal{M}}(\boldsymbol{x}_i(t), \hat{\boldsymbol{x}}_i(t))^2 \leq \frac{C_1(\mathcal{M},\mathcal{I},T)^2}{N}\sum_{i=1}^N \|\boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t)\|_{\mathbb{R}^{d'}}^2 \\ &\leq \frac{2C_1(\mathcal{M},\mathcal{I},T)^2 T\exp(64T^2S_0^2)}{N}\sum_{i=1}^N \int_{s=0}^t \left\| \dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\hat{\phi},ii',s}\right\|_{\mathbb{R}^{d'}}^2 ds. \\ &= \frac{2C_1(\mathcal{M},\mathcal{I},T)^2 T\exp(64T^2S_0^2)}{N}\sum_{i=1}^N \int_{s=0}^t \left\| \dot{\boldsymbol{x}}_i(s) - \frac{1}{N}\sum_{i'=1}^N F^{\mathcal{M}}_{\hat{\phi},ii',s}\right\|_{T_{\boldsymbol{x}_i(s)}\mathcal{M}}^2 ds \\ &= 2C_1(\mathcal{M},\mathcal{I},T)^2 T\exp(64T^2S_0^2)\int_{s=0}^t \left\| \dot{\boldsymbol{X}}_s - \boldsymbol{f}^{\mathrm{c}}_{\hat{\phi}}(\boldsymbol{X}_s)\right\|_{T_{\boldsymbol{X}_s}\mathcal{M}^N}^2 ds \end{aligned}$$

Letting

$$C(\mathcal{M},T) := \inf_{\text{all isometric embeddings } \mathcal{I}} C_1(\mathcal{M},\mathcal{I},T)^2,$$

and choosing an isometric embedding $\mathcal{I}$ which gives a value at most twice the infimum, we obtain

$$d_{\mathcal{M}}(\boldsymbol{X}_t, \hat{\boldsymbol{X}}_t)^2 \leq 4TC(\mathcal{M},T)\exp(64T^2S_0^2)\int_{s=0}^t \left\| \dot{\boldsymbol{X}}_s - \boldsymbol{f}^{\mathrm{c}}_{\hat{\phi}}(\boldsymbol{X}_s)\right\|_{T_{\boldsymbol{X}}\mathcal{M}^N}^2 ds.$$

Now, take ϕ to be the true interaction kernel, and $\hat{\phi}$ the estimator of ϕ by our learning approach, by Prop. 4.1 we have that

$$\frac{1}{T}\int_{t=0}^T \left\| \dot{\boldsymbol{X}}_s - \boldsymbol{f}^{\mathrm{c}}_{\hat{\phi}}(\boldsymbol{X}_s)\right\|_{T_{\boldsymbol{X}}\mathcal{M}^N}^2 dt \leq \left\| \phi(\cdot)\cdot - \hat{\phi}(\cdot)\cdot\right\|_{L^2(\rho_{T,\mathcal{M}})}^2.$$

Together with (4.8.11), recalling that $\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$ and $\boldsymbol{X}_0 \sim \mu^{\boldsymbol{x}}$, we have the desired result that

$$\mathbb{E}_{\boldsymbol{X}_0\sim\mu^{\boldsymbol{x}}}\left[d_{\mathrm{traj},\mathcal{M}}(\boldsymbol{X}_{[0,T]}, \hat{\boldsymbol{X}}_{[0,T]})^2\right] \leq 4T^2C(\mathcal{M},T)\exp(64T^2S_0^2)\mathbb{E}_{\boldsymbol{X}_0\sim\mu^{\boldsymbol{x}}}\left\| \phi(\cdot)\cdot - \hat{\phi}(\cdot)\cdot\right\|_{L^2(\rho_{T,\mathcal{M}})}^2.$$

□

4.9 Numerical Implementations

If the trajectory data, $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M}$, is given by the user, we use the following geometry-based algorithm to find the minimizer of (4.8.2). First, we construct a finite dimensional subspace of the hypothesis space, i.e. $\mathcal{H}_M \subset \mathcal{H}$, where $\mathcal{H}_M$ with dimension $d(\mathcal{H}_M) = n = n(M) \approx \mathcal{O}(M^{\frac{1}{3}})$ is a space of clamped B-spline functions[1] supported on $[R_{\min}^{\text{obs}}, R_{\max}^{\text{obs}}]$ with $R_{\min}^{\text{obs}}/R_{\max}^{\text{obs}}$ being the minimum/maximum interaction radius computed from the observation data. Hence the test functions can be expressed as linear combination of the basis functions of $\mathcal{H}_M$, i.e., $\varphi(r) = \sum_{\eta=1}^{n} \alpha_\eta \psi_\eta(r)$ with $\{\psi_\eta\}_{\eta=1}^n$ being a basis for $\mathcal{H}_M$. Next, we use either a local chart $\mathcal{U} : \mathcal{M} \to \mathbb{R}^d$ or a natural embedding $\mathcal{I} : \mathcal{M} \to \mathbb{R}^{d'}$, such that $\boldsymbol{x}_i \in \mathcal{M}$ can be expressed using either local coordinates in $\mathbb{R}^d$ (as in the $\mathbb{PD}$ case) or global coordinates in $\mathbb{R}^{d'}$ (as in the $\mathbb{S}^2$ case). The computation of $\langle \cdot, \cdot, \rangle\, g(\boldsymbol{x})$ will be based on the choice of the local chart, or on the embedding, accordingly. Then, we define a basis matrix, $\Psi^m \in (T_{\boldsymbol{X}_{t_1}^m}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}_{t_L}^m}\mathcal{M}^N)^n$, whose columns are

$$\Psi^m(:,\eta) = \Psi_\eta^m = \frac{1}{\sqrt{N}} \begin{bmatrix} \boldsymbol{f}_{\psi_\eta}^{\text{c}}(\boldsymbol{X}_{t_1}^m) \\ \vdots \\ \boldsymbol{f}_{\psi_\eta}^{\text{c}}(\boldsymbol{X}_{t_L}^m) \end{bmatrix} \in T_{\boldsymbol{X}_{t_1}^m}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}_{t_L}^m}\mathcal{M}^N,$$

recall

$$\boldsymbol{f}_\varphi^{\text{c}}(\boldsymbol{X}_t) = \begin{bmatrix} \vdots \\ \frac{1}{N}\sum_{i'=1}^N \varphi(d_{\mathcal{M}}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)))\boldsymbol{w}(\boldsymbol{x}_i(t), \boldsymbol{x}_{i'}(t)) \\ \vdots \end{bmatrix} \in T_{\boldsymbol{X}_t}\mathcal{M}^N.$$

[1] Other type of basis functions can be considered, such as piecewise polynomials, Fourier, etc.

Next, we define the derivative vector, $\vec{d}^m \in T_{\boldsymbol{X}_{t_1}^m}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}_{t_L}^m}\mathcal{M}^N$, as follows,

$$\vec{d}^m = \frac{1}{\sqrt{N}} \begin{bmatrix} \dot{\boldsymbol{X}}_{t_1}^m \\ \vdots \\ \dot{\boldsymbol{X}}_{t_L}^m \end{bmatrix}.$$

Then, we define the learning matrix $A_M \in \mathbb{R}^{n\times n}$ as follows

$$A_M(\eta,\eta') = \frac{1}{LM}\sum_{m=1}^{m} \langle \Psi_\eta^m, \Psi_{\eta'}^m, \rangle \, G, \quad \text{for } \eta,\eta' = 1,\dots,n.$$

Here the inner product $\langle \cdot,\cdot, \rangle \, G$ on $\Psi_\eta^m \in T_{\boldsymbol{X}_{t_1}^m}\mathcal{M}^N \times \cdots \times T_{\boldsymbol{X}_{t_L}^m}\mathcal{M}^N$ is defined as

$$\langle \Psi_\eta^m, \Psi_{\eta'}^m, \rangle \, G = \sum_{l=1}^{L} \langle \boldsymbol{f}_{\psi_\eta}^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m), \boldsymbol{f}_{\psi_{\eta'}}^{\mathrm{c}}(\boldsymbol{X}_{t_l}^m), \rangle \, g^{\mathcal{M}^N}(\boldsymbol{X}_l^m).$$

Next for the learning right hand side, $\vec{b}_M \in \mathbb{R}^{n\times 1}$, we have

$$\vec{b}_M(\eta) = \frac{1}{LM}\sum_{m=1}^{m} \langle \vec{d}, \Psi_\eta^m, \rangle \, G, \quad \text{for } \eta = 1,\dots,n$$

Therefore, the minimization of (4.8.2) over $\mathcal{H}_M$ can be rewritten as

$$A_M\vec{\alpha} = \vec{b}_M, \quad \vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^{n\times 1}.$$

A_M is symmetric positive definite (guaranteed by the geometric coercivity condition), hence we can solve the linear system to obtain $\widehat{\vec{\alpha}}$, and assemble

$$\hat{\phi}(r) = \sum_{\eta=1}^{n} \hat{\alpha}_\eta \psi_\eta(r).$$

In order to produce unique solution of (4.8.1) using $\hat{\phi}$, we smooth out $\hat{\phi}$ for the evolution of the dynamics.

If the trajectory data is not given, we will generate it using a Geometric Numerical Integrator, which is a fourth order Backward Differentiation Formula (BDF) of fixed time step size h combined with a projection scheme. For details see [67]. Once a reasonable evolution of the dynamics is obtained, we observe it at $0 = t_1 < \ldots < t_L = T$ to obtain a set of trajectory data, and use it as training data to input to the learning algorithm. The observation times do not need to be aligned with the numerical integration times, i.e. where numerical solution of $\{\boldsymbol{x}_i^m(t), \dot{\boldsymbol{x}}_i^m(t)\}_{i,m=1}^{N,M}$ is obtained at $\{t_{l'}\}_{l'=1}^{L'}$ (except for $t_1 = 0$ and $t_{L'} = T$). When t_l does not land on one of the numerical integration time points, a continuous extension method is used to interpolate the numerical solution at t_l.

4.10 Numerical Experiments

We consider three prototypical first order dynamics, Opinion Dynamics (OD), Lennard-Jones Dynamics (LJD), and Predator-Swarm dynamics (PS1), on two different manifolds, the $2D$ sphere ($\mathbb{S}^2$ centered at the origin with radius $\frac{5}{\pi}$) and the Poincaré disk ($\mathbb{PD}$, unit disk centered at the origin, with the hyperbolic metric). The two prototypical manifolds are chosen because $\mathbb{S}^2$ and $\mathbb{PD}$ are model spaces with constant positive and negative curvature, respectively. We conduct extensive experiments on the aforementioned six different scenarios to demonstrate the performance of our learning approach for dynamics evolving on manifolds. We report the results in terms of function estimation errors and trajectory estimation errors, and discuss in detail the learning performance of the estimators.

The setup of the numerical experiments is as follows. We generate a set of M_ρ different initial conditions, and evolve the various dynamics of N agents for

$t \in [0,T]$ using a Geometric Numerical Integrator with a uniform time step h (for details see section 4.9); then we observe each dynamics at equidistant times, i.e. $0 = t_1 < \ldots < t_L = T$, to obtain a set of trajectory data, $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i,l,m=1}^{N,L,M_\rho}$, to approximate the "true" probability distribution $\rho_{T,\mathcal{M}}^L$. From this set of pre-generated trajectory data, we randomly choose a subset of $M \ll M_\rho$ of them to be used as training data for the learning simulation. The hypothesis space where the estimator is learned is generated as a set of n first-degree clamped B-spline basis functions built on a uniform partition of the learning interval $[R_{\min}^{\text{obs}}, R_{\max}^{\text{obs}}]$, with $R_{\min}^{\text{obs}}$ and $R_{\max}^{\text{obs}}$ being the minimum and maximum interaction radii computed from the training and trajectory data, respectively. Once an estimator, denoted as $\hat{\phi}$, is obtained, we report the estimation error, $\phi(\cdot)\cdot - \hat{\phi}(\cdot)\cdot$, using

$$\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{\text{Rel}.L^2(\rho_{T,\mathcal{M}})} := \frac{\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}{\|\phi(\cdot)\cdot\|_{L^2(\rho_{T,\mathcal{M}})}}; \tag{4.10.1}$$

and the trajectory estimation error

$$d_{\text{trj}}(\boldsymbol{X}_{[0,T]}^m, \hat{\boldsymbol{X}}_{[0,T]}^m)^2 := \sup_{t\in[0,T]} \frac{\sum_i d_{\mathcal{M}}(\boldsymbol{x}_i^m(t), \hat{\boldsymbol{x}}_i^m(t))^2}{N} \tag{4.10.2}$$

between, the true and estimated dynamics, evolved using ϕ or $\hat{\phi}$ with the same initial conditions for $t \in [0,T]$ respectively, and observed at the same observation times $0 = t_1 < \ldots < t_L = T$, over both the training initial conditions and another set of M randomly chosen initial conditions. Moreover, the above learning procedure is run 10 times independently in order to generate empirical error bars. We will report the errors in the form of mean $\pm$ std. Visual comparisons of ϕ versus $\hat{\phi}$, and $\boldsymbol{X}$ versus $\hat{\boldsymbol{X}}$ will be shown, and discussions of learning results will be presented in each subsection.

Table 4.5 shows the values of the common parameters shared by all six experiments.

M_ρ	N	L	M	Num. of Learning Trials	$R_{\mathcal{M}}$ on $\mathbb{S}^2$	$R_{\mathcal{M}}$ on $\mathbb{PD}$
3000	20	500	500	10	5	∞

Table 4.5: Values of the parameters shared by the six experiments

Moreover, section 4.7 shows the details on how to calculate the geodesic direction and the Riemannian distance between any two points on $\mathbb{S}^2$ and $\mathbb{PD}$. The distribution of the initial conditions, $\mu^{\boldsymbol{x}}$, is given as follows: uniform on $\mathcal{M} = \mathbb{S}^2$; whereas uniform on an open ball (centered at origin with radius r_0) for the $\mathbb{PD}$ case with r_0 given as follows.

$$r_0 = \left(2 + \frac{1}{\cosh(5) - 1} - \sqrt{\frac{4}{\cosh(5) - 1} + \frac{1}{(\cosh(5) - 1)^2}}\right)/2.$$

This radius is used so that the maximum distance between any pair of agents on the Poincaré disk is 5. PS1 will have different setup for the initial conditions, which will be discussed in section 4.10.4.

4.10.1 Computing Platform

We use a computing workstation with an AMD Ryzen 9 3900X CPU (which has 12 computing cores), and available 128 GB memory, running CentOS 7, provided and managed by Prisma Analytics, Inc. . All 6 experiments are ran in the MATLAB (R2020a) environment with parallel mode enabled and a parallel pool of 12 workers. Such parallel mode is used in each experiment for the computation of $\rho^L_{T,\mathcal{M}}$, learning, and trajectory error estimation. Detailed report of the running time for the experiments is provided in the result section of each experiment.

4.10.2 Opinion Dynamics

We first choose opinion dynamics, which is used to model simple interactions of opinions [6, 139] as well as choreography [26]. We consider the generalization of

this dynamics to take place on two different manifolds: the $2D$ sphere ($\mathbb{S}^2$) and the Poincaré disk ($\mathbb{PD}$). We consider the interaction kernel

$$\phi(r) := \begin{cases} 1, & 0 \le r < \frac{1}{\sqrt{2}} - 0.01 \\ a_1 r^3 + b_1 r^2 + c_1 r + d_1, & \frac{1}{\sqrt{2}} - 0.01 \le r < \frac{1}{\sqrt{2}} \\ 0.1, & \frac{1}{\sqrt{2}} \le r < 0.99 \\ a_2 r^3 + b_2 r^2 + c_2 r + d_2, & 0.99 \le r < 1 \\ 0, & \text{otherwise} \end{cases}$$

The parameters, i.e. $(a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2)$, are chosen so that $\phi \in C^1([0,1])$. Table 4.6 shows the values of the parameters needed for the learning simulation.

$n_{\mathbb{S}^2}$	$n_{\mathbb{PD}}$	T	h
51	69	10	0.01

Table 4.6: Test Parameters for OD.

Results for the $\mathbb{S}^2$ case: Fig. 4.3 shows the comparison between ϕ and its estimator $\hat{\phi}$ learned from the trajectory data.

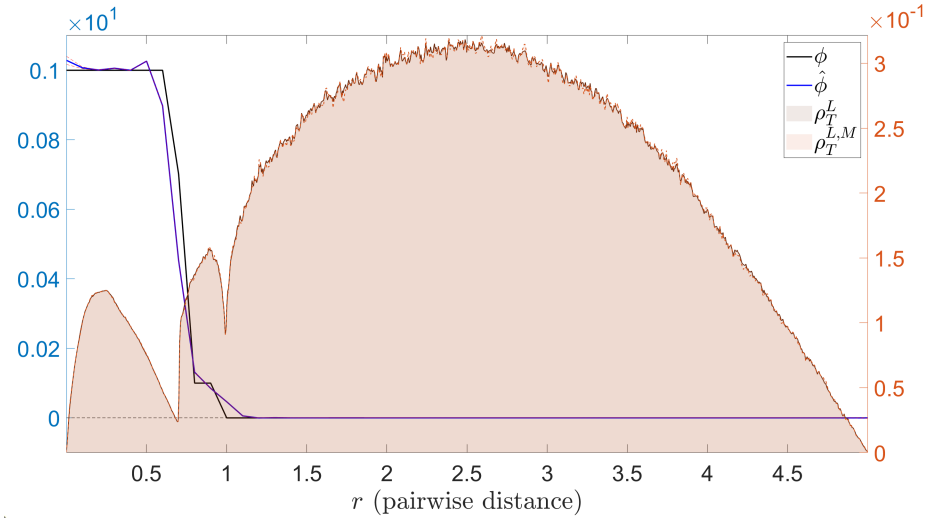


Figure 4.3: (OD on $\mathbb{S}^2$) Comparison of ϕ and $\hat{\phi}$, with the relative error being $1.894 \cdot 10^{-1} \pm 3.1 \cdot 10^{-4}$ (calculated using (4.10.1)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in red dotted lines. Shown in the background is the comparison of the approximate $\rho^L_{T,\mathcal{M}}$ versus the empirical $\rho^{L,M}_{T,\mathcal{M}}$.

As it is shown in Fig. 4.3, the estimator is able to capture the compact support of the ϕ from the trajectory data. Fig. 4.4 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

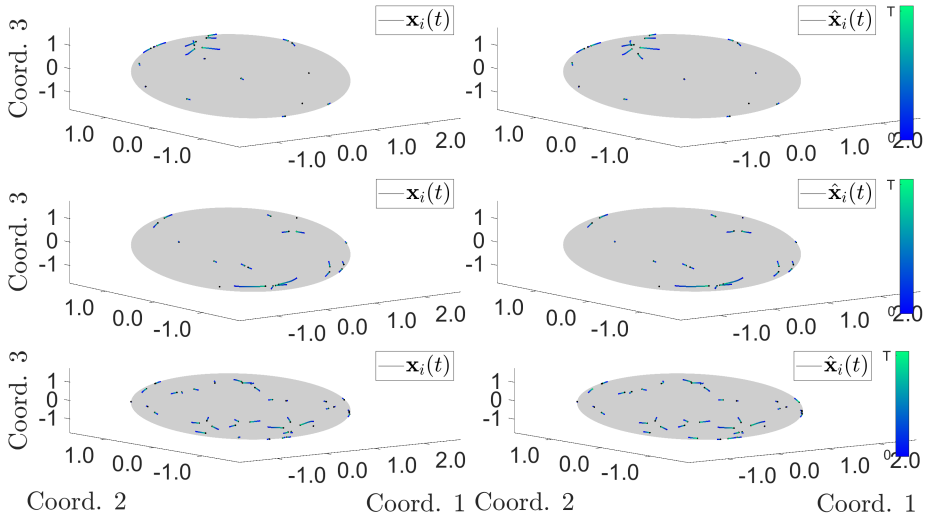


Figure 4.4: (OD on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\hat{\phi}$), with the errors reported in table 4.7. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

A quantitative comparison of the trajectory estimation errors is shown in Table 4.7.

	$[0, T]$
mean_{IC}: Training ICs	$8.8 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$
std_{IC}: Training ICs	$5.9 \cdot 10^{-2} \pm 1.5 \cdot 10^{-3}$
mean_{IC}: Random ICs	$9.0 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$
std_{IC}: Random ICs	$6.0 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$

Table 4.7: (OD on $\mathbb{S}^2$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). mean_{IC} and std_{IC} are the mean and standard deviation of the trajectory errors calculated using (4.10.2).

We also report the condition number and the smallest eigenvalue of the learning matrix A to indirectly verify the geometric coercivity condition in table 4.8.

Condition Number	$1.8 \cdot 10^5 \pm 1.4 \cdot 10^4$
Smallest Eigenvalue	$1.09 \cdot 10^{-7} \pm 9.0 \cdot 10^{-9}$

Table 4.8: (OD on $\mathbb{S}^2$) Information from the learning matrix A.

It took $1.41 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $4.76 \cdot 10^4$ seconds to run 10 learning simulations, with $1.44 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.44 \cdot 10^2 \pm 3.1$ seconds to run one estimation), and $4.61 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $4.61 \cdot 10^3 \pm 20.0$ seconds to run one set of trajectory error estimation).

Results for the $\mathbb{PD}$ case: Fig. 4.5 shows the comparison between the C^1 version of ϕ and its estimator $\hat{\phi}$ learned from the trajectory data.

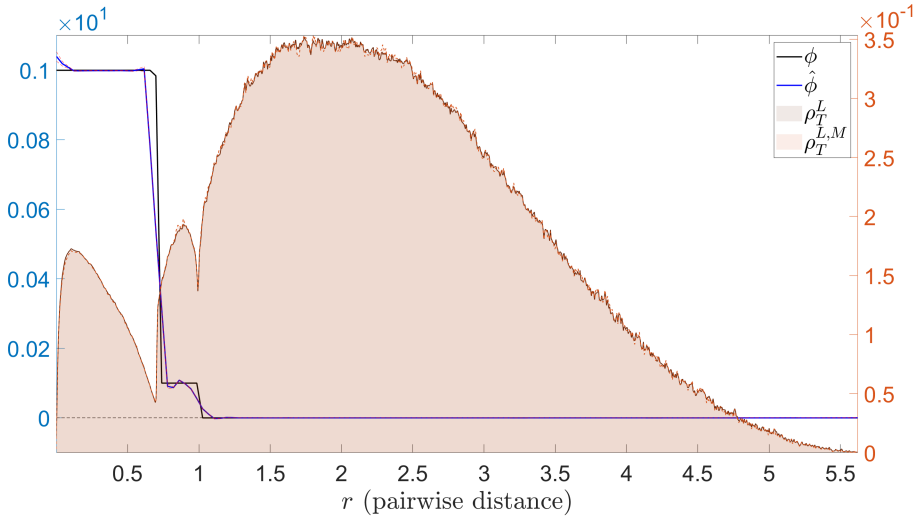


Figure 4.5: (OD on $\mathbb{PD}$) Comparison of ϕ and $\hat{\phi}$, with the relative error being $2.114 \cdot 10^{-1} \pm 5.0 \cdot 10^{-4}$ (calculated using (4.10.1)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in red dotted lines. Shown in the background is the comparison of the approximate $\rho^L_{T,\mathcal{M}}$ versus the empirical $\rho^{L,M}_{T,\mathcal{M}}$.

As it is shown in Fig. 4.5, the estimator is able to capture the compact support of the ϕ from the trajectory data. Fig. 4.6 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

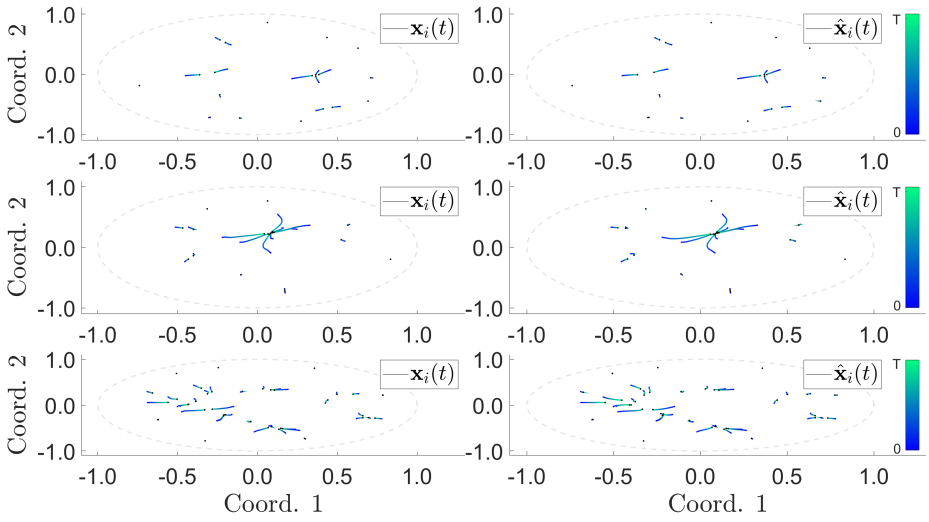


Figure 4.6: (OD on $\mathbb{PD}$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\hat{\phi}$), with the errors reported in table 4.9. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

As shown in Fig. 4.5, around $r = \frac{1}{\sqrt{2}}$, the estimator $\hat{\phi}$ produces values bigger than that from ϕ, leading to stronger influence, hence the merging of cluster happening in the predicted trajectories in the second row of Fig. 4.6. As demonstrated by the average prediction error on trajectories, this is a relatively rare event, occurring for only certain initial conditions. A quantitative comparison of the trajectory estimation errors is shown in Table 4.9.

	$[0, T]$
mean_{IC}: Training ICs	$2.53 \cdot 10^{-1} \pm 7.2 \cdot 10^{-3}$
std_{IC}: Training ICs	$1.90 \cdot 10^{-1} \pm 6.5 \cdot 10^{-3}$
mean_{IC}: Random ICs	$2.55 \cdot 10^{-1} \pm 9.7 \cdot 10^{-3}$
std_{IC}: Random ICs	$1.89 \cdot 10^{-1} \pm 5.9 \cdot 10^{-3}$

Table 4.9: (OD on $\mathbb{PD}$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). mean_{IC} and std_{IC} are the mean and standard deviation of the trajectory errors calculated using (4.10.2).

We also report the condition number and the smallest eigenvalue of the learning matrix A to indirectly verify the geometric coercivity condition in table 4.10.

Condition Number	$4.9 \cdot 10^5 \pm 1.5 \cdot 10^4$
Smallest Eigenvalue	$5.3 \cdot 10^{-6} \pm 1.2 \cdot 10^{-7}$

Table 4.10: (OD on $\mathbb{PD}$) Information from the learning matrix A.

It took $1.33 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $4.06 \cdot 10^4$ seconds to run 10 learning simulations, with $1.23 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.23 \cdot 10^2 \pm 1.1$ seconds to run one estimation), and $3.93 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $3.93 \cdot 10^3 \pm 82.1$ seconds to run one set of trajectory error estimation).

4.10.3 Lennard-Jones Dynamics

The second first-order model considered here is induced from a special energy functional, the so-called Lennard-Jones energy potential. This first-order model, the Lennard-Jones Dynamics (LJD), is a simplified version of the second-order dynamics used in molecular dynamics. The energy function, U_{LJ}, is given by

$$U_{\text{LJ}}(r) := 4\varepsilon\Big(\Big(\frac{\sigma}{r}\Big)^{12} - \Big(\frac{\sigma}{r}\Big)^{6}\Big).$$

Here ε is the depth of the potential well, σ is the distance when U is zero, and r is the distance between any pair of agents. We set $\varepsilon = 10$ and $\sigma = 1$. The corresponding interaction kernel ϕ, derived from this potential, is

$$\phi_{\text{LJ}}(r) := \frac{U'_{\text{LJ}}(r)}{r} = 24\frac{\varepsilon}{\sigma^2}\Big(\Big(\frac{\sigma}{r}\Big)^{8} - 2\Big(\frac{\sigma}{r}\Big)^{14}\Big).$$

We shall use a slightly modified version of ϕ_{LJ}:

$$\phi(r) := \begin{cases} \phi_{\text{LJ}}(1) - \phi'_{\text{LJ}}(1)/4, & 0 \le r < \frac{1}{2} \\ \phi'_{\text{LJ}}(1)r^2 - \phi'_{\text{LJ}}(1)r + \phi_{\text{LJ}}(1), & \frac{1}{2} \le r < 1 \\ \phi_{\text{LJ}}(r), & 1 \le r < 0.99R_{\mathcal{M}} \\ a_3 r^3 + b_3 r^2 + c_3 r + d_3, & 0.99R_{\mathcal{M}} \le r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \le r. \end{cases}$$

The parameters, (a_3, b_3, c_3, d_3), are chosen so that $\phi \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi(r) = \phi_{\text{LJ}}(r)$ for $r \ge 1$. Table 4.11 shows the values of the parameters needed for the learning simulation.

$n_{\mathbb{S}^2}$	$n_{\mathbb{PD}}$	T	h
51	69	10^{-3}	10^{-6}

Table 4.11: Test Parameters for LJD.

Results for the $\mathbb{S}^2$ case: Fig. 4.7 shows the comparison between ϕ and its estimator $\hat{\phi}$ learned from the trajectory data.

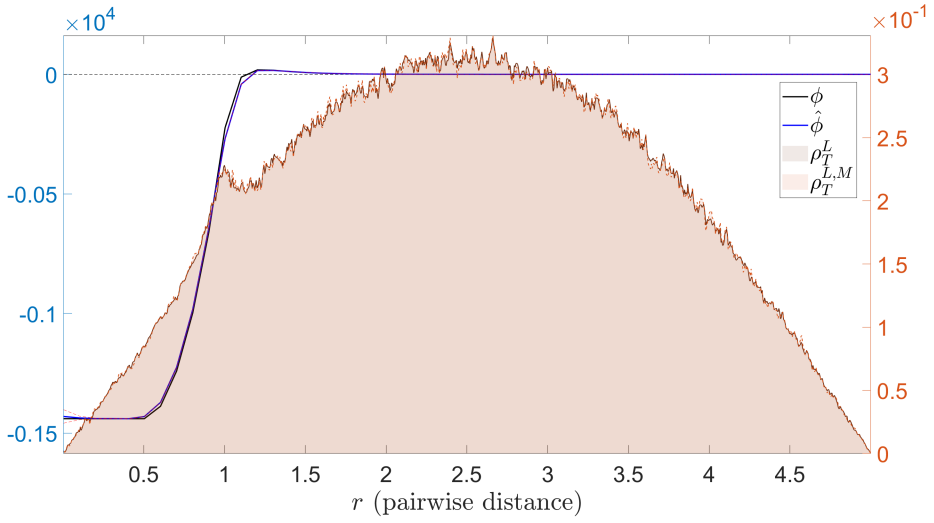


Figure 4.7: (LJD on $\mathbb{S}^2$) Comparison of ϕ and $\hat{\phi}$, with the relative error being $3.65 \cdot 10^{-2} \pm 2.7 \cdot 10^{-4}$ (calculated using (4.10.1)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in blue dotted lines. Shown in the background is the comparison of the approximate ρ_T^L versus the empirical $\rho_T^{L,M}$.

Fig. 4.8 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

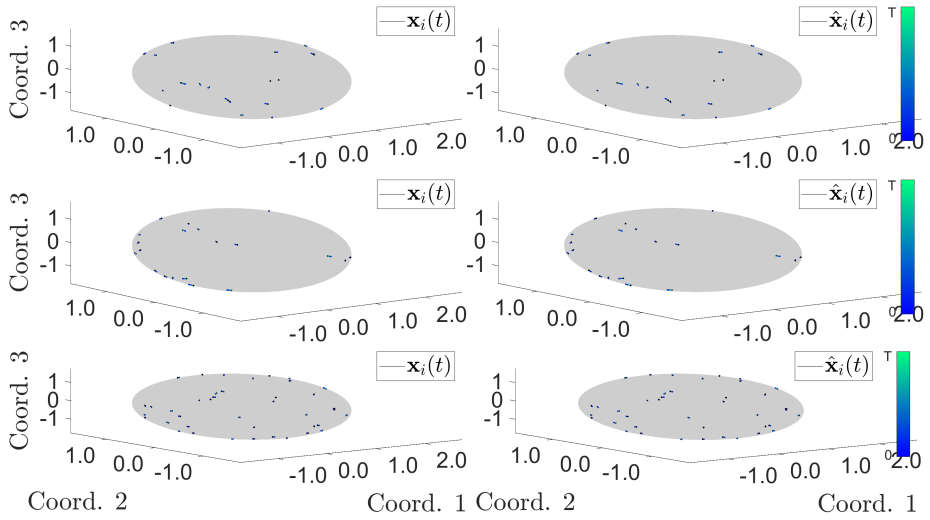


Figure 4.8: (LJD on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\hat{\phi}$), with the errors reported in table 4.12. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

A quantitative comparison of the trajectory estimation errors is shown in Table 4.12.

	$[0, T]$
mean_{IC}: Training ICs	$2.88 \cdot 10^{-3} \pm 2.5 \cdot 10^{-5}$
std_{IC}: Training ICs	$6.1 \cdot 10^{-4} \pm 1.8 \cdot 10^{-5}$
mean_{IC}: Random ICs	$2.88 \cdot 10^{-3} \pm 3.2 \cdot 10^{-5}$
std_{IC}: Random ICs	$6.0 \cdot 10^{-4} \pm 1.8 \cdot 10^{-5}$

Table 4.12: (LJD on $\mathbb{S}^2$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). The trajectory estimation errors is calculated using (4.10.1).

We also report the condition number and the smallest eigenvalue of the learning matrix A to indirectly verify the geometric coercivity condition in table 4.13.

Condition Number	$6 \cdot 10^5 \pm 1.5 \cdot 10^5$
Smallest Eigenvalue	$2.4 \cdot 10^{-8} \pm 6.2 \cdot 10^{-9}$

Table 4.13: (LJD on $\mathbb{S}^2$) Information from the learning matrix A.

It took $2.43 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $7.14 \cdot 10^4$ seconds to run 10 learning simulations, with $1.72 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.72 \cdot 10^2 \pm 2.5$ seconds to run one estimation), and $6.96 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $6.96 \cdot 10^3 \pm 35.9$ seconds to run one set of trajectory error estimation).

Results for the $\mathbb{PD}$ case: Fig. 4.9 shows the comparison between ϕ and its estimator $\hat{\phi}$ learned from the trajectory data.

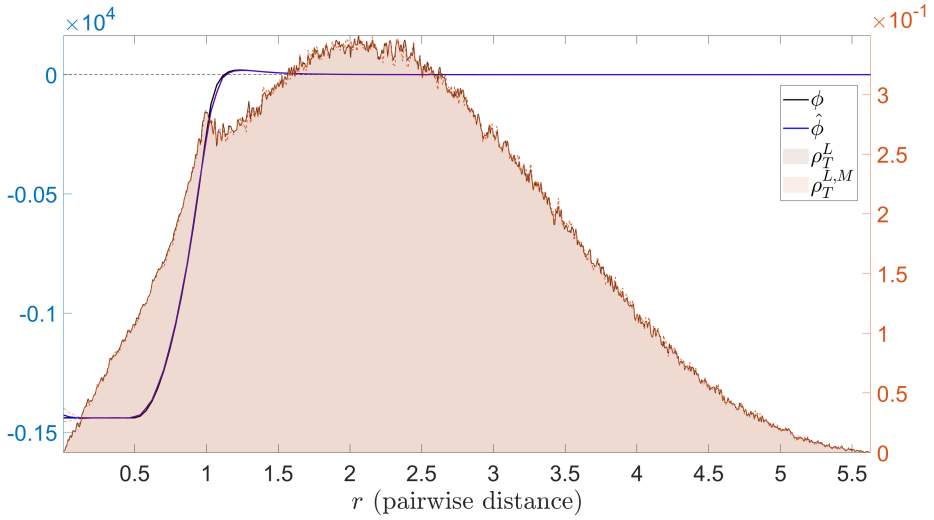


Figure 4.9: (LJD on $\mathbb{PD}$) Comparison of ϕ and $\hat{\phi}$, with the relative error being $2.52 \cdot 10^{-2} \pm 3.6 \cdot 10^{-4}$ (calculated using (4.10.1)). The true interaction kernel is shown in black solid line, whereas the mean estimated interaction kernel is shown in blue solid line with its confidence interval shown in blue dotted lines. Shown in the background is the comparison of the approximate ρ^L_T versus the empirical $\rho^{L,M}_T$.

Fig. 4.10 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

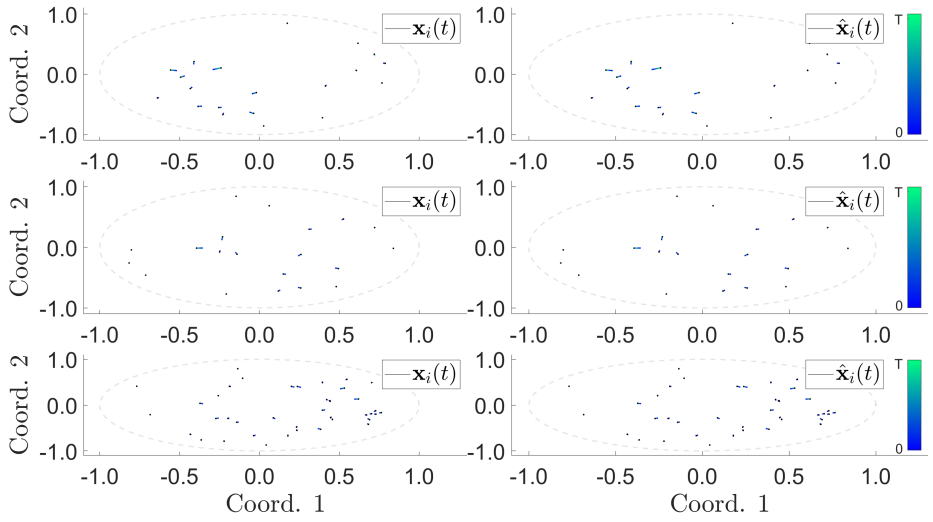


Figure 4.10: (LJD on $\mathbb{PD}$) Comparison of $\boldsymbol{X}$ (generated by ϕ) and $\hat{\boldsymbol{X}}$ (generated by $\hat{\phi}$), with the errors reported in table 4.14. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue (at $t = 0$) to light green (at $t = T$).

A quantitative comparison of the trajectory estimation errors is shown in Table 4.14.

	$[0, T]$
mean_{IC}: Training ICs	$2.27 \cdot 10^{-3} \pm 4.0 \cdot 10^{-5}$
std_{IC}: Training ICs	$5.6 \cdot 10^{-4} \pm 1.7 \cdot 10^{-5}$
mean_{IC}: Random ICs	$2.28 \cdot 10^{-3} \pm 3.8 \cdot 10^{-5}$
std_{IC}: Random ICs	$5.6 \cdot 10^{-4} \pm 1.6 \cdot 10^{-5}$

Table 4.14: (LJD on $\mathbb{PD}$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). mean_{IC} and std_{IC} are the mean and standard deviation of the trajectory errors calculated using (4.10.2).

We also report the condition number and the smallest eigenvalue of the learning matrix A to indirectly verify the geometric coercivity condition in table 4.15.

Condition Number	$6 \cdot 10^{6} \pm 1.9 \cdot 10^{6}$
Smallest Eigenvalue	$1.7 \cdot 10^{-8} \pm 6.6 \cdot 10^{-9}$

Table 4.15: (LJD on $\mathbb{PD}$) Information from the learning matrix A.

It took $1.51 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $6.23 \cdot 10^4$ seconds to run 10 learning simulations, with $1.20 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.20 \cdot 10^2 \pm 9.4$ seconds to run one estimation), and $6.10 \cdot 10^4$ seconds spent on computing the trajectory error estimates (on average, it took $6 \cdot 10^3 \pm 1.3 \cdot 10^3$ seconds to run one set of trajectory error estimation).

4.10.4 Predator-Swarm Dynamics

The third first-order model considered here is a heterogeneous agent system, which is used to model interactions between multiple types of animals [32, 103] or agents (need ref.). The learning theory presented in this work is described for homogeneous agent systems, but the theory and the corresponding algorithms extend naturally to heterogeneous agent systems in a manner analogous to [90, 96].

We consider here a system of a single predator versus a group of preys, namely the Predator-Swarm Dynamics (PS1), discussed in [32]. The preys are in type 1, and the single predator is in type 2. We have multiple interaction kernels, depending on the types of agents in each interacting pair: $\phi_{kk'}$ defines the influence of agents in type k' on agents in type k, for $k, k' = 1, 2$. The interaction kernels are given as follows.

$$\phi_{11}(r) := \begin{cases} \frac{2}{0.01^3}(r - 0.01) + (1 - \frac{1}{0.01^2}) & 0 < r \leq 0.01 \\ 1 - \frac{1}{r^2} & 0.01 < r \leq 0.99 R_{\mathcal{M}} \\ a_{1,1} r^3 + b_{1,1} r^2 + c_{1,1} r + d_{1,1}, & 0.99 R_{\mathcal{M}} \leq r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \leq r \end{cases}$$

The parameters, $(a_{1,1}, b_{1,1}, c_{1,1}, d_{1,1})$, are chosen so that $\phi_{11}(r) \in C^1([0, R_{\mathcal{M}}])$ when

$R_{\mathcal{M}} < \infty$; otherwise $\phi_{11}(r) = 1 - \frac{1}{r^2}$ for $r \geq 0.01$;

$$\phi_{12}(r) := \begin{cases} \frac{4}{0.01^3}(r - 0.01) + \frac{-2}{0.01^2}) & 0 < r \leq 0.01 \\ \frac{-2}{r^2} & 0.01 < r \leq 0.99 R_{\mathcal{M}} \\ a_{1,2} r^3 + b_{1,2} r^2 + c_{1,2} r + d_{1,2}, & 0.99 R_{\mathcal{M}} \leq r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \leq r \end{cases}$$

The parameters, $(a_{1,2}, b_{1,2}, c_{1,2}, d_{1,2})$, are chosen so that $\phi_{12}(r) \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi_{12}(r) = \frac{-2}{r^2}$ for $r \geq 0.01$;

$$\phi_{21}(r) := \begin{cases} \frac{-10.5}{0.01^4}(r - 0.01) + \frac{3.5}{0.01^3}) & 0 < r \leq 0.01 \\ \frac{3.5}{r^3} & 0.01 < r \leq 0.99 R_{\mathcal{M}} \\ a_{2,1} r^3 + b_{2,1} r^2 + c_{2,1} r + d_{2,1}, & 0.99 R_{\mathcal{M}} \leq r < R_{\mathcal{M}} \\ 0, & R_{\mathcal{M}} \leq r \end{cases}$$

The parameters, $(a_{2,1}, b_{2,1}, c_{2,1}, d_{2,1})$, are chosen so that $\phi_{21}(r) \in C^1([0, R_{\mathcal{M}}])$ when $R_{\mathcal{M}} < \infty$; otherwise $\phi_{21}(r) = \frac{3.5}{r^3}$ for $r \geq 0.01$; then $\phi_{22} \equiv 0$, since there is only one predator. We set $T = 0.5$ and $h = 10^{-4}$ for the two $PS1$ models.

Results for the $\mathbb{S}^2$ case: In order to produce more interesting interactions, we choose the distribution of the initial condition to be as follows. The setting will start from $\mathbb{R}^2$ first. The position of the predator is randomly chosen uniformly within a circular disk of radius 0.1 centered at the origin of $\mathbb{R}^2$. The remaining $N - 1$ agents will be prey and chosen uniformly at random within an annulus of radii 0.3 and 0.8, centered at the origin. Then these positions will mapped through a stereographic projection (where the origin of $\mathbb{R}^2$ is the south pole of $\mathbb{S}^2$) back to $\mathbb{S}^2$. When back on $\mathbb{S}^2$, the position of the predator is moved via parallel transport to a random location on $\mathbb{S}^2$, and the rest of the preys are moved using the same map, so that the relative position between each pair of agents is not changed.

Table 4.16 shows the number of basis functions, namely $n_{kk'}$'s, for each estimator $\hat{\phi}_{kk'}$ for $k, k' = 1, 2$, and their corresponding degrees, $p_{k,k'}$'s, for the Clamped B-spline basis.

$n_{1,1}$	$n_{1,2}$	$n_{2,1}$	$n_{2,2}$
50	37	37	1
$p_{1,1}$	$p_{1,2}$	$p_{2,1}$	$p_{2,2}$
1	1	1	0

Table 4.16: (PS1 on $\mathbb{S}^2$) Number of basis functions.

Fig. 4.13 shows the comparison between $\phi_{kk'}$ and its estimators $\hat{\phi}_{kk'}$ learned from the trajectory data.

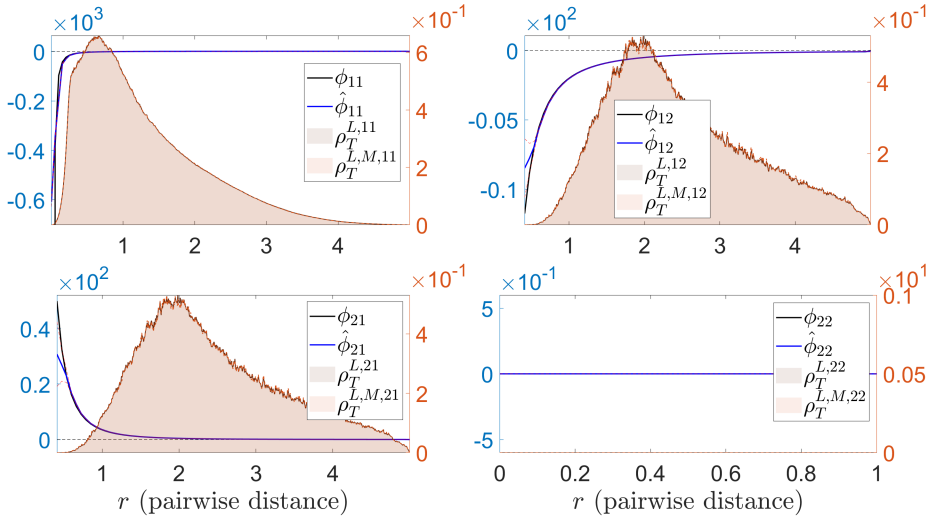


Figure 4.11: (PS1 on $\mathbb{S}^2$) Comparison of $\phi_{kk'}$ and $\hat{\phi}_{k,k'}$, with the relative errors shown in table 4.21. The true interaction kernels are shown in black solid line, whereas the mean estimated interaction kernel are shown in blue solid line with their corresponding confidence intervals shown in blue dotted lines. Shown in the background is the comparison of the approximate $\rho_T^{L,kk'}$ versus the empirical $\rho_T^{L,M,kk'}$. Notice that $\rho_T^{L,12}/\rho_T^{L,M,12}$ and $\rho_T^{L,12}/\rho_T^{L,M,21}$ are the same distributions.

$\text{Err}_{1,1}$	$\text{Err}_{1,2}$	$\text{Err}_{2,1}$	$\text{Err}_{2,2}$
$2.98 \cdot 10^{-1} \pm 5.9 \cdot 10^{-3}$	$8.4 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$	$2.5 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$	0

Table 4.17: (PS1 on $\mathbb{S}^2$) Relative estimation errors calculated using (4.10.1).

Fig. 4.12 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

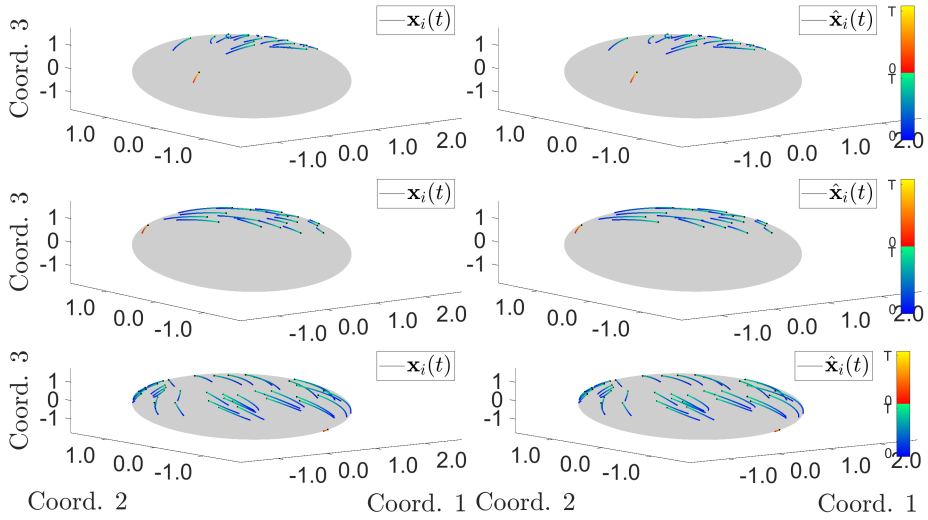


Figure 4.12: (PS1 on $\mathbb{S}^2$) Comparison of $\boldsymbol{X}$ (generated by $\phi_{k,k'}$'s) and $\hat{\boldsymbol{X}}$ (generated by $\hat{\phi}_{k,k'}$'s), with the errors reported in table 4.18. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue/bright red (at $t = 0$) to light green/light yellow (at $t = T$). The blue/green combination is assigned to the preys; whereas the red/yellow comb for the predator.

A quantitative comparison of the trajectory estimation errors is shown in Table 4.22.

	$[0, T]$
mean_{IC}: Training ICs	$2.36 \cdot 10^{-2} \pm 9.8 \cdot 10^{-4}$
std_{IC}: Training ICs	$1.9 \cdot 10^{-2} \pm 1.5 \cdot 10^{-4}$
mean_{IC}: Random ICs	$2.40 \cdot 10^{-2} \pm 8.1 \cdot 10^{-4}$
std_{IC}: Random ICs	$2.3 \cdot 10^{-3} \pm 6.1 \cdot 10^{-3}$

Table 4.18: (PS1 on $\mathbb{S}^2$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). mean_{IC} and std_{IC} are the mean and standard deviation of the trajectory errors calculated using (4.10.2).

We also report the condition number and the smallest eigenvalue of the learning matrix A to indirectly verify the geometric coercivity condition in table 4.23.

Condition Number for A_1	$2.2 \cdot 10^7 \pm 1.8 \cdot 10^6$
Smallest Eigenvalue for A_1	$1.28 \cdot 10^{-8} \pm 8.5 \cdot 10^{-10}$
Condition Number for A_2	$2.9 \cdot 10^5 \pm 2.2 \cdot 10^5$
Smallest Eigenvalue for A_2	$9 \cdot 10^{-7} \pm 5.7 \cdot 10^{-7}$

Table 4.19: (PS1 on $\mathbb{S}^2$) Information from the learning matrix A_k's.

The matrix A_1 is used to obtain the estimators, $\hat{\phi}_{1,1}$ and $\hat{\phi}_{1,2}$; whereas A_2 is used to obtain $\hat{\phi}_{2,1}$ and $\hat{\phi}_{2,2}$. Since there is one single predator, we set $\hat{\phi}_{2,2}$ to zero. It took $9.77 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $4.01 \cdot 10^5$ seconds to run 10 learning simulations, with $1.66 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.66 \cdot 10^2 \pm 4.6$ seconds to run one estimation), and $4.05 \cdot 10^5$ seconds spent on computing the trajectory error estimates (on average, it took $4.0 \cdot 10^4 \pm 7.1 \cdot 10^3$ seconds to run one set of trajectory error estimation).

Results for the $\mathbb{PD}$ case: In order to produce more interesting interactions, we choose the distribution of the initial condition to be as follows: the predator is randomly placed in a circle centered at the origin with radius r_1, given as follows

$$r_0 = \left(2 + \frac{1}{\cosh(0.5) - 1} - \sqrt{\frac{4}{\cosh(0.5) - 1} + \frac{1}{(\cosh(0.5) - 1)^2}}\right)/2,$$

so that the agents are at most 0.5 distance away from each other; then the group of preys (Swarm) will be randomly and uniformly placed on an annulus centered at the origin with radii,(R_1, r_1), given as follows

$$r_1 = \left(2 + \frac{1}{\cosh(1) - 1} - \sqrt{\frac{4}{\cosh(1) - 1} + \frac{1}{(\cosh(1) - 1)^2}}\right)/2$$

and

$$R_1 = \left(2 + \frac{1}{\cosh(2) - 1} - \sqrt{\frac{4}{\cosh(2) - 1} + \frac{1}{(\cosh(2) - 1)^2}}\right)/2;$$

so that the group of preys are surrounding the single predator. Table 4.20 shows the number of basis functions, namely $n_{kk'}$'s, for each estimator $\hat{\phi}_{kk'}$ for $k, k' = 1, 2$, and

their corresponding degrees, $p_{k,k'}$'s, for the Clamped B-spline basis.

$n_{1,1}$	$n_{1,2}$	$n_{2,1}$	$n_{2,2}$
68	43	43	1
$p_{1,1}$	$p_{1,2}$	$p_{2,1}$	$p_{2,2}$
1	1	1	0

Table 4.20: (PS1 on $\mathbb{PD}$) Number of basis functions.

Fig. 4.13 shows the comparison between $\phi_{kk'}$ and its estimators $\hat{\phi}_{kk'}$ learned from the trajectory data.

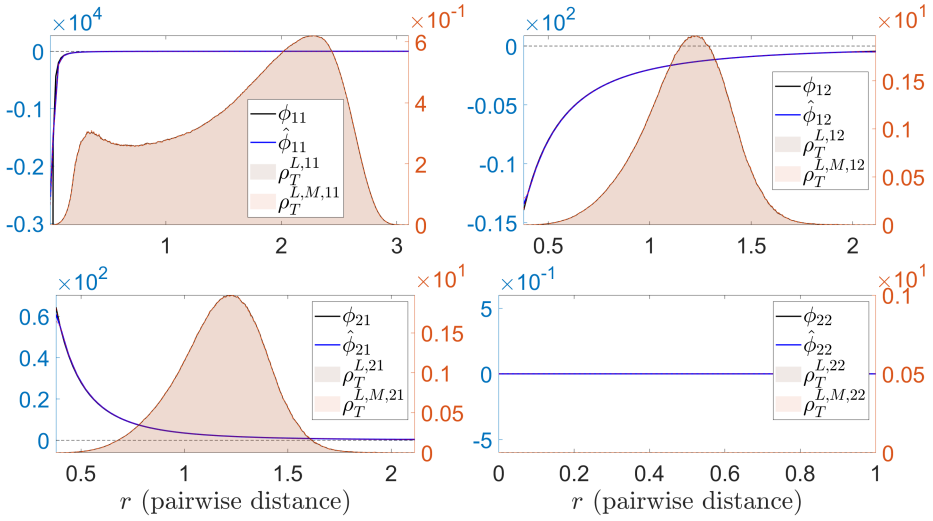


Figure 4.13: (PS1 on $\mathbb{PD}$) Comparison of $\phi_{kk'}$ and $\hat{\phi}_{k,k'}$, with the relative errors shown in table 4.21. The true interaction kernels are shown in black solid line, whereas the mean estimated interaction kernel are shown in blue solid line with their corresponding confidence intervals shown in blue dotted lines. Shown in the background is the comparison of the approximate $\rho_T^{L,kk'}$ versus the empirical $\rho_T^{L,M,kk'}$. Notice that $\rho_T^{L,12}/\rho_T^{L,M,12}$ and $\rho_T^{L,12}/\rho_T^{L,M,21}$ are the same distributions.

$\text{Err}_{1,1}$	$\text{Err}_{1,2}$	$\text{Err}_{2,1}$	$\text{Err}_{2,2}$
$9.0 \cdot 10^{-2} \pm 2.6 \cdot 10^{-3}$	$1.34 \cdot 10^{-3} \pm 8.8 \cdot 10^{-5}$	$3.6 \cdot 10^{-3} \pm 2.4 \cdot 10^{-4}$	0

Table 4.21: (PS1 on $\mathbb{PD}$) Relative estimation errors calculated using (4.10.1).

Fig. 4.14 shows the comparison of the trajectory data between the true dynamics and estimated dynamics.

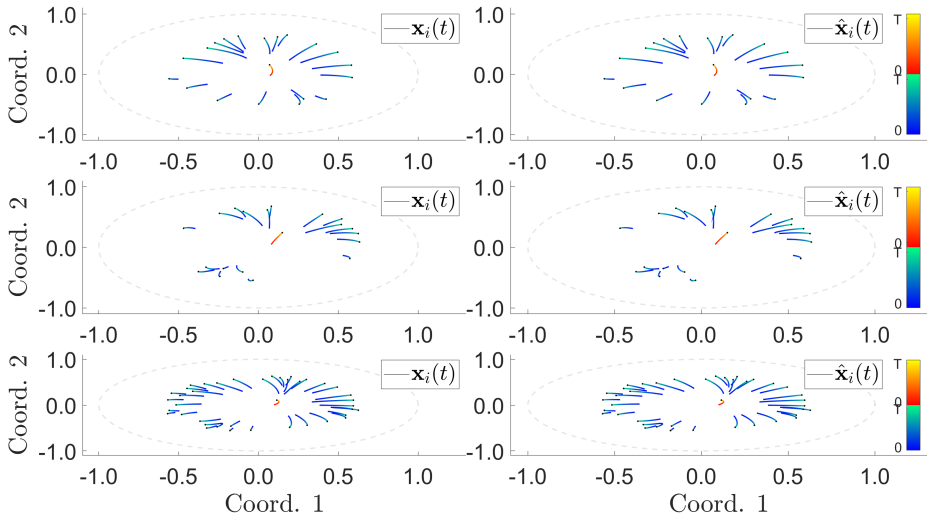


Figure 4.14: (PS1 on $\mathbb{PD}$) Comparison of $\boldsymbol{X}$ (generated by $\phi_{k,k'}$'s) and $\hat{\boldsymbol{X}}$ (generated by $\hat{\phi}_{k,k'}$'s), with the errors reported in table 4.22. **Top**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from an initial condition taken from the training data. **Middle**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a randomly chosen initial condition. **Bottom**: $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are generated from a new initial condition with bigger $N = 40$. The color of the trajectory indicates the flow of time, from deep blue/bright red (at $t = 0$) to light green/light yellow (at $t = T$). The blue/green combination is assigned to the preys; whereas the red/yellow comb for the predator.

A quantitative comparison of the trajectory estimation errors is shown in Table 4.22.

	$[0, T]$
mean_{IC}: Training ICs	$4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$
std_{IC}: Training ICs	$2.3 \cdot 10^{-3} \pm 3.0 \cdot 10^{-4}$
mean_{IC}: Random ICs	$4.8 \cdot 10^{-3} \pm 1.2 \cdot 10^{-4}$
std_{IC}: Random ICs	$2.5 \cdot 10^{-3} \pm 3.9 \cdot 10^{-3}$

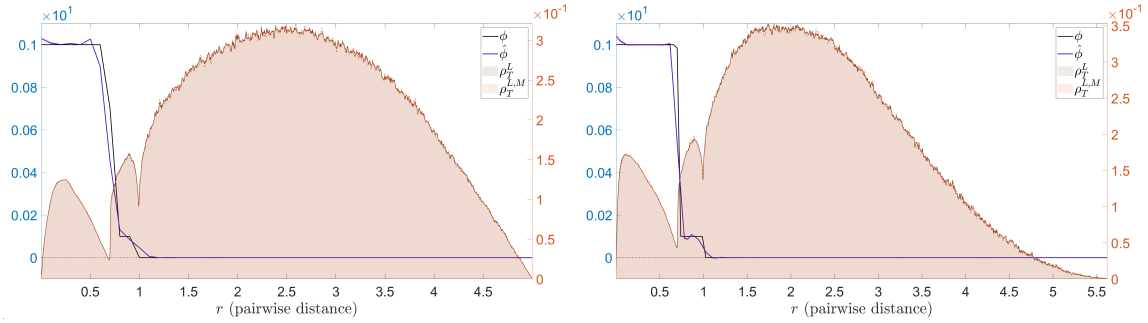
Table 4.22: (PS1 on $\mathbb{PD}$) trajectory estimation errors: Initial Conditions (ICs) used in the training set (first two rows), new ICs randomly drawn from $\mu^{\boldsymbol{x}}$ (second set of two rows). mean_{IC} and std_{IC} are the mean and standard deviation of the trajectory errors calculated using (4.10.2).

We also report the condition number and the smallest eigenvalue of the learning matrix A to indirectly verify the geometric coercivity condition in table 4.23.

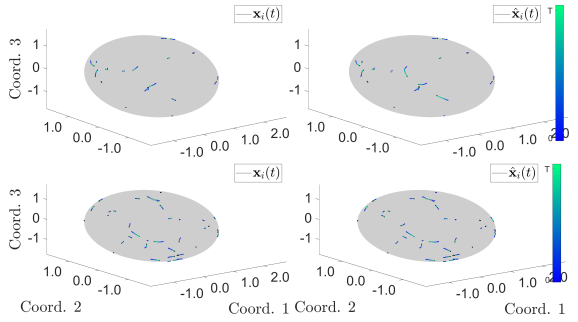
Condition Number for A_1	$2.3 \cdot 10^9 \pm 4.7 \cdot 10^8$
Smallest Eigenvalue for A_1	$7 \cdot 10^{-11} \pm 1.7 \cdot 10^{-11}$
Condition Number for A_2	$5 \cdot 10^5 \pm 3.1 \cdot 10^5$
Smallest Eigenvalue for A_2	$4 \cdot 10^{-8} \pm 2.9 \cdot 10^{-8}$

Table 4.23: (PS1 on $\mathbb{PD}$) Information from the learning matrix A_k's.

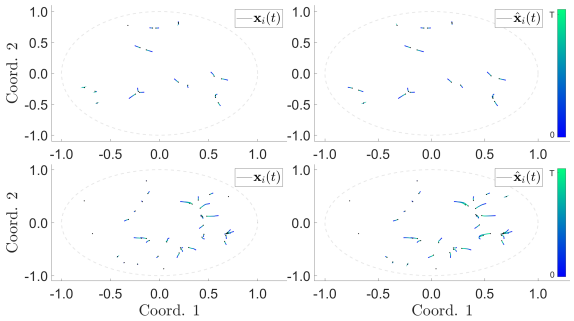
The matrix A_1 is used to obtain the estimators, $\hat{\phi}_{1,1}$ and $\hat{\phi}_{1,2}$; whereas A_2 is used to obtain $\hat{\phi}_{2,1}$ and $\hat{\phi}_{2,2}$. Since there is one single predator, we set $\hat{\phi}_{2,2}$ to zero. It took $7.37 \cdot 10^4$ seconds to generate $\rho^L_{T,\mathcal{M}}$ and $2.49 \cdot 10^5$ seconds to run 10 learning simulations, with $1.25 \cdot 10^3$ seconds spent on learning the estimated interactions (on average, it took $1.25 \cdot 10^2 \pm 1.5$ seconds to run one estimation), and $2.48 \cdot 10^5$ seconds spent on computing the trajectory error estimates (on average, it took $2.48 \cdot 10^4 \pm 2.3 \cdot 10^2$ seconds to run one set of trajectory error estimation).



((a)) OD on $\mathbb{S}^2$ ((b)) OD on $\mathbb{PD}$

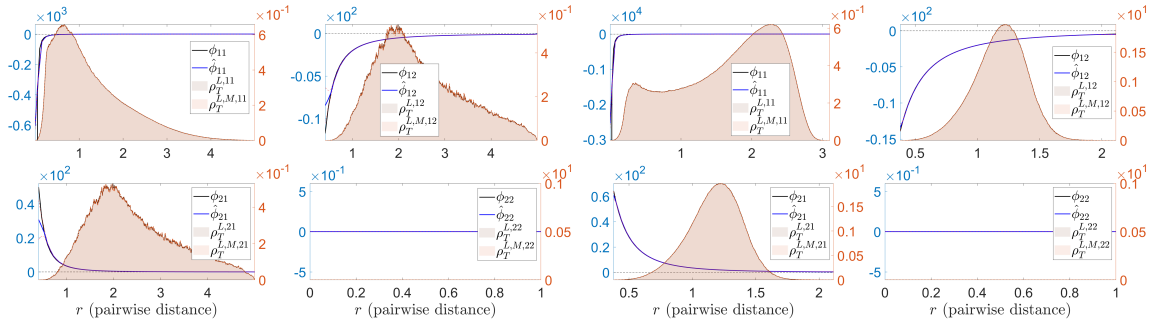


((c)) OD on $\mathbb{S}^2$

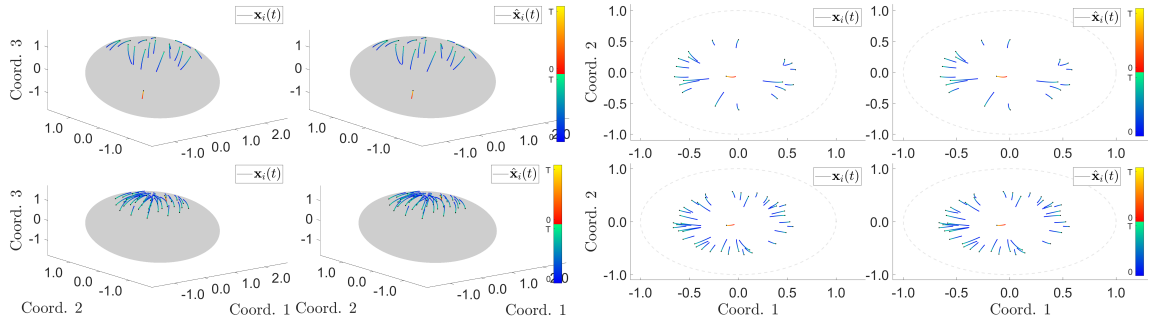


((d)) OD on $\mathbb{PD}$

Figure 4.1: Top: comparison of ϕ and $\hat{\phi}$. The true interaction kernel is shown with a black solid line, whereas the mean estimated interaction kernel is shown with a blue solid line with its confidence interval shown in red dotted lines. Shown in the background is the comparison of the approximate $\rho^L_{T,\mathcal{M}}$ versus the empirical $\rho^{L,M}_{T,\mathcal{M}}$. **Bottom**: comparison of trajectories $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$. The trajectories $\boldsymbol{X}_{[0,T]}$'s generated by the true interaction kernel ϕ; whereas $\hat{\boldsymbol{X}}_{[0,T]}$'s are trajectories generated by the estimator $\hat{\phi}$, with the same initial conditions. In the first row, trajectories are started from a randomly chosen initial condition. In the second row, trajectories are generated for a new system, with $N = 40$ agents. The colors along the trajectories indicate time, from deep blue (at $t = 0$) to light green (at $t = T$).



((a)) PS1 on $\mathbb{S}^2$ ((b)) PS1 on $\mathbb{PD}$



((c)) PS1 on $\mathbb{S}^2$ ((d)) PS1 on $\mathbb{PD}$

Figure 4.2: Top: comparison of $\phi_{k,k'}$ and $\hat{\phi}_{k,k'}$. The true interaction kernels are shown with a black solid line, whereas the mean estimated interaction kernels are shown with a blue solid line with their confidence intervals shown in red dotted lines. Shown in the background is the comparison of the approximate $\rho_{T,\mathcal{M}}^{L,kk'}$ versus the empirical $\rho_{T,\mathcal{M}}^{L,M,kk'}$. Notice that $\rho_T^{L,12}/\rho_T^{L,M,12}$ and $\rho_T^{L,12}/\rho_T^{L,M,21}$ are the same distributions. **Bottom**: comparison of trajectories $\boldsymbol{X}_{[0,T]}$ and $\hat{\boldsymbol{X}}_{[0,T]}$. The trajectories $\boldsymbol{X}_{[0,T]}$'s generated by the true interaction kernel $\phi_{k,k'}$; whereas $\hat{\boldsymbol{X}}_{[0,T]}$'s are trajectories generated by the estimator $\hat{\phi}_{k,k'}$, with the same initial conditions. In the first row, trajectories are started from a randomly chosen initial condition. In the second row, trajectories are generated for a new system, with $N = 40$ agents. The colors along the trajectories indicate time, from deep blue/bright red (at $t = 0$) to light green/light yellow (at $t = T$). The blue/green combo is assigned to tge preys; whereas red/yellow combo to the predator.

Chapter 5

Numerical Experiments

5.1 Introduction

Our approach is developed based on a collective dynamics framework, derived from classical Lagrangian mechanics. We begin with a second order system in the form

$$m_i \ddot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^{N} \frac{1}{N} \phi(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|)(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)), \tag{5.1.1}$$

for $i = 1, \ldots, N$. Here, $\phi : \mathbb{R}^+ \to \mathbb{R}$, is known as an **interaction kernel**. The problem of inferring the interaction kernel from observed trajectories, in a non-parametric fashion, was considered in [19], and extended in multiple directions (both theoretical and of practical relevance) in [89, 146, 96, 93] (see the SI for a detailed discussion).

In this work, we consider the problem of inferring the interaction laws of celestial bodies in the Solar System from trajectory data, with minimal a priori knowledge about their form; in particular we assume no knowledge of geometric properties of the trajectories (e.g. elliptical, closed, etc.), of masses of the celestial bodies (in fact, not even the concept of mass), and no assumption on the form of the interaction kernels (e.g. inverse powers of pairwise distance). We are particularly interested in discovering effective laws of gravitation that best explain empirical data (both from

observations and simulations) and compare these effective laws to celebrated models from Physics.We use trajectory data from the Jet Propulsion Laboratory's (JPL) development ephemerides. We compare the performance of our models to the JPL data, as well as to two important simulated systems: one based on Newton's gravity model, and the other based on the Einstein-Infeld-Hoffman equations. We discover that our models can provide superior performance over the other two simulated systems in terms of trajectory error estimation, preserving the geometric properties (period/aphelion/perihelion) of the trajectories, and reproducing the highly sensitive and localized perihelion precession rates of three prototypical bodies: Mars (observation of its orbits led to classical Newtonian gravity), Mercury (where the general relativity effect is prominent), the Moon (where gravity alone cannot provide a full explanation of the precession), see table 5.1 for details.

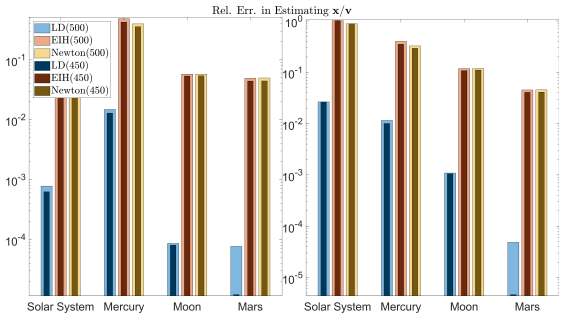


Figure 5.1: Comparison of relative errors in estimating position/velocity (using (5.4.2) and (5.4.3)) from three different dynamics (LD/EIH/Newton) compared to the JPL's observation data for the full solar system, Mercury, the Moon, and Mars over 450 and 500 year trajectories. The errors over 450 years have smaller width and darker color, and are laid on top of the errors over 500 years, which have greater width and a lighter color. Different colors correspond to different dynamics: dark/light blue for LD, dark/light red for EIH, and dark/light yellow for Newton. The learned dynamics demonstrates high accuracy in terms of trajectory error in all cases.

5.2 Results

We present the most significant results from our machine-learning procedure for celestial dynamics (labeled as LD for "Learned Dynamics") compared to the JPL's

observation data (JPL), simulated EIH system (EIH), and simulated Newton's system (Newton). The errors in estimating position and velocity are shown in Figure 5.1. LD demonstrates superior performance in terms of estimating the trajectories, on both training (a period of 450 years) and testing (a subsequent period of 50 years) data. Table 5.1 shows the perihelion precession rates of Mercury, the Moon, and Mars from the four different models considered, estimated over 450 years of trajectory data.

	JPL	LD	EIH	Newton
Mercury	576.58	567.28	569.48	533.35
Moon	$3.43 \cdot 10^7$	$3.49 \cdot 10^7$	$3.07 \cdot 10^7$	$2.80 \cdot 10^7$
Mars	$1.52 \cdot 10^3$	$1.50 \cdot 10^3$	$1.52 \cdot 10^3$	$1.52 \cdot 10^3$

Table 5.1: Perihelion precession rate (PPR) estimation for 3 different celestial bodies from 4 different models. The algorithm for calculating the PPR is presented in the SI. Our learned dynamics estimates accurately the precession rate of all 3 celestial bodies, substantially better than Newton on Mercury, where it is nearly as accurate as EIH, and is the most accurate at the complicated precession dynamics of the moon.

Our estimation procedure is able to capture the essence of the PPR, which is the most sensitive, localized, and representative of the characteristics of the dynamics for 3 prototypical celestial bodies, with the least amount of assumptions made on the observation data. In Figure 5.2, we show the estimated dynamics evolved using our estimators for 500 years using a symplectic integrator described in the Learning Results section.

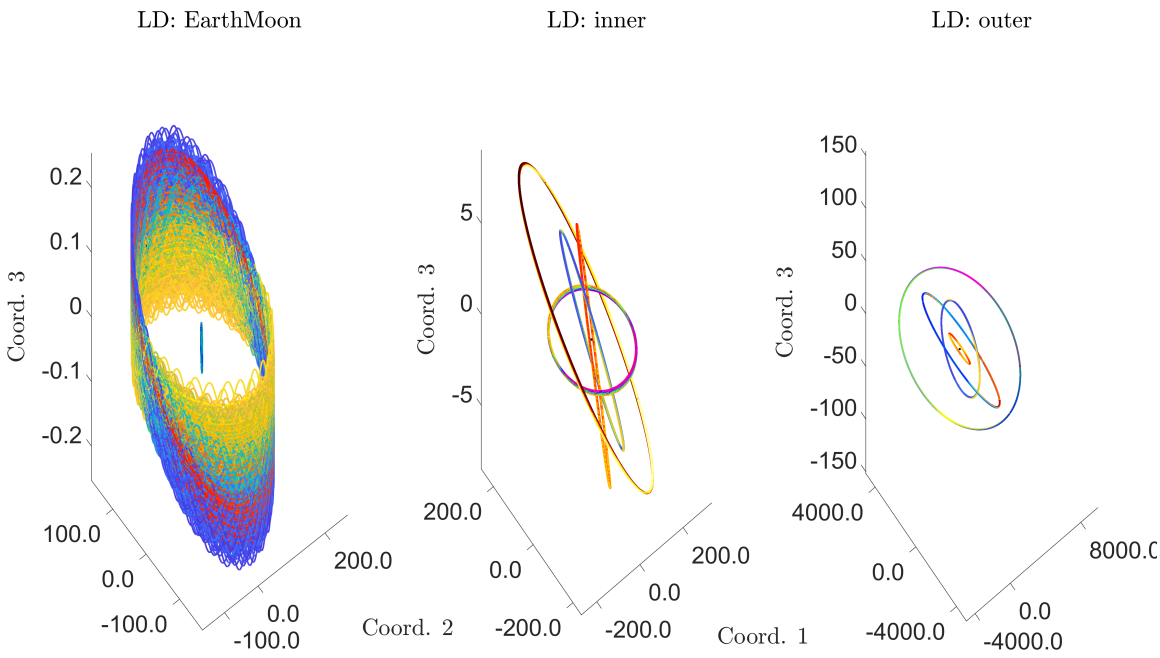


Figure 5.2: Trajectories of the Solar System from the learned estimators (LD) evolved over 500 years with the initial position/velocity taken at year 1500 from the JPL data. Left: Earth-Moon-Sun system; **Middle**: Inner Solar system; **Right**: Outer Solar system. Since we are maintaining 10^{-3} relative accuracy at estimating the positions for the celestial bodies in our Solar system, we will not show the trajectory plots of the true data from JPL.

5.3 Model Description

We make the following assumptions in order to simplify the discussion of our models: *i*) we will be working with absolute time and space, which is valid in the low energy/low velocity setting of the Solar system; *ii*) the gravitational mass is the the same as the inertial mass in the equations of motion; *iii*) the gravitational effect acts instantaneously at any distance. Improvements incorporating relativistic principles into our models is a focus of future work. In our machine learning based approach we assume that the gravitational force depends only on pairwise distance, which is consistent with translation and rotation-invariance; we do not assume any analytic expression for the interactions, and we do not assume knowledge of the concept mass, nor of how it affects interactions. We work with each celestial body as if it were of its

own "type", with its own type of interaction with other celestial bodies.

Using the framework of Lagrangian Mechanics, we consider the Lagrangian, $\mathcal{L}(t)$, of a closed Solar system of N celestial bodies, each identified with its center of mass:

$$\mathcal{L}(t) = \sum_{i=1}^{N} \frac{1}{2} m_i \left\| \boldsymbol{v}_i(t) \right\|^2 - \frac{1}{2} \sum_{i,i'=1}^{N} U_{i,i'}(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|).$$

Here $\boldsymbol{x}_i, \boldsymbol{v}_i \in \mathbb{R}^d$ ($d = 2$ or 3) are the position or velocity of the i^{th} celestial body, $\|\cdot\|$ is the usual Euclidean norm, and $U_{i,i'} : \mathbb{R}^+ \to \mathbb{R}$ is a potential energy depending only on pairwise distance, and it is parameterized by the unknown masses of celestial body i and i' in an unknown, possibly non-linear way. We further assume that $U_{i,i'} = U_{i',i}$ and $U_{i,i} \equiv 0$. Then, via the Lagrange equation, $\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_i}\right) = \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i}$, we arrive at the equation of motion for the i^{th} celestial body:

$$m_i \dot{\boldsymbol{v}}_i(t) = \sum_{i'=1}^{N} (U'_{i,i'}(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|)) \cdot \frac{\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)}{\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|}.$$

Remark 5.3.1. *In the case of Newton's gravitational potential, we have* $U_{i,i'}(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|) = \frac{-Gm_i m_{i'}}{\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|}$. *Newton came to the conclusion of this particular form based on Kepler's laws and the assumption that the gravitational force should have a "*$\frac{1}{r^p}$*" form with* $p = 2$ *being the only solution for closed elliptical orbits.*

Simplifying, we obtain the following equations of motion for the i^{th} celestial body ($i = 1, \ldots, N$),

$$\dot{\boldsymbol{v}}_i(t) = \sum_{i'=1, i' \neq i}^{N} \phi_{i,i'}(r_{i,i'}(t)) \boldsymbol{r}_{i,i'}(t). \tag{5.3.1}$$

Here $\boldsymbol{r}_{i,i'}(t) \coloneqq \boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)$, $r_{i,i'}(t) \coloneqq \|\boldsymbol{r}_{i,i'}(t)\|$, and $[\phi_{i,i'}]_{i,i'=1}^{N}$ is a set of **interaction kernels**, with each $\phi_{i,i'}$ induced by an unknown $\phi_{i,i'}(r) = \frac{U'_{i,i'}(r)}{m_i r}$. Hence, $\phi_{i,i'}$ contains hidden information about the masses of celestial body i and i', namely, m_i and $m_{i'}$.

5.4 Learning Framework

Our learning approach makes no prior assumptions on the trajectory (such as it being elliptical), uses no knowledge of the masses of the celestial bodies, and has no prior knowledge of the particular functional forms of the interaction kernels – the learning algorithm is merely given a set of discrete time trajectory data for the celestial bodies in our Solar system, namely, $\{\boldsymbol{x}_i(t_l), \boldsymbol{v}_i(t_l), \dot{\boldsymbol{v}}_i(t_l)\}_{i,l=1}^{N,L}$ for $T_0 = t_1 < \cdots < t_L = T$. The estimation procedure for the interaction kernels for models given by (5.3.1) is based on the variational approaches introduced in [19, 89, 146, 96, 93]. We introduce vectorized notation as follows: let $\boldsymbol{X}_{t_l}$ (resp. $\boldsymbol{V}_{t_l}$) be the column vector obtained by concatenating the column vectors $(\boldsymbol{x}_i(t_l))_{i=1}^N$ (resp.: $(\boldsymbol{v}_i(t_l))_{i=1}^N$), and

$$\mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}_{t_l}) \coloneqq \begin{bmatrix} \vdots \\ \sum_{i'=1}^N \varphi_{i,i'}(r_{i,i'}(t_l))\boldsymbol{r}_{i,i'}(t_l) \\ \vdots \end{bmatrix}.$$

Here $\boldsymbol{\varphi} \coloneqq [\varphi_{i,i'}]_{i,i'=1}^N$. Notice $\boldsymbol{X}_{t_l}, \boldsymbol{V}_{t_l}, \mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}_{t_l}) \in \mathbb{R}^D$, with $D \coloneqq Nd$. We define the $\|\cdot\|_{\mathcal{S}}$ on $\mathbb{R}^D$ as $\|\boldsymbol{X}\|_{\mathcal{S}}^2 \coloneqq \sum_{i=1}^N \|\boldsymbol{x}_i\|^2$.

5.4.1 Non-parametric Learning of Interaction Kernels

To simplify the discussion, we take equispaced time points, i.e. $t_l - t_{l-1} = t_{l+1} - t_l$ for $l = 2, \ldots, L-1$; however equispacing is not mandatory for our algorithm. We find a set of estimated interaction kernels $\widehat{\boldsymbol{\phi}} \coloneqq [\hat{\phi}_{i,i'}]_{i,i'=1}^N$ by minimizing the following L^2 error functional,

$$\mathcal{E}_L(\boldsymbol{\varphi}) \coloneqq \frac{1}{L}\sum_{l=1}^{L} \left\| \ddot{\boldsymbol{X}}_{t_l} - \mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{X}_{t_l}) \right\|_{\mathcal{S}}^2. \tag{5.4.1}$$

Here, $\boldsymbol{\varphi} = [\varphi_{i,i'}]_{i,i'=1}^N$ with each $\varphi_{i,i'} \in \mathcal{H}_{i,i'}$ a compact (in the L^∞-norm) and convex subset of $L^2([R_{i,i'}^{\min}, R_{i,i'}^{\max}])$. Let $\boldsymbol{\mathcal{H}} \coloneqq \oplus_{i,i'=1}^N \mathcal{H}_{i,i'}$ and $\widehat{\boldsymbol{\phi}} \coloneqq \underset{\boldsymbol{\varphi}\in\mathcal{H}}{\arg\min}\ \mathcal{E}_L(\boldsymbol{\varphi})$. The

convergence of $\widehat{\boldsymbol{\phi}}$ to the true interaction kernels, in the more restrictive setting that does include unknown masses or other parameters, is studied in [96]: one of the major takeaways of that analysis is that even if the estimation problem is for a system in D dimension, upon choosing suitable hypothesis spaces, with dimension suitably growing with the number of observations (as in nonparametric statistics methods), one can achieve a learning rate that only depends on the number of variables in the interaction kernel, which in this case is 1 (pairwise distance).

5.4.2 Performance Measures

We consider two other types of performance measures, both of which depend on the difference in observed planetary motion and estimated planetary motion – which are generated by evolving the dynamical system using the learned estimators or known physical laws (EIH or Newton) starting at the same initial conditions (IC) as the observation data. Let $\boldsymbol{X}_t$ be the observed positions of the N-body system for $t \in [T_0, T]$, and $\hat{\boldsymbol{X}}_t$ be the estimated positions evolved from the same IC as the observed system with the estimated interaction kernels or known equations of motion defined by EIH or Newton over $t \in [T_0, T]$, then consider

$$\mathrm{Err}_1 := \frac{\max_{t\in[T_0,T]} \left\| \boldsymbol{X}_t - \hat{\boldsymbol{X}}_t \right\|_{\mathcal{S}}}{\max_{t\in[T_0,T]} \|\boldsymbol{X}_t\|_{\mathcal{S}}}. \tag{5.4.2}$$

However, by considering the system as a whole, errors in an individual celestial body's trajectory with smaller positions could be obscured. To avoid this, we consider and report a second type of error,

$$\mathrm{Err}_{2,i} := \frac{\max_{t\in[T_0,T]} \|\boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t)\|}{\max_{t\in[T_0,T]} \|\boldsymbol{x}_i(t)\|} \tag{5.4.3}$$

We use similar formulas for computing estimation errors in velocities $\boldsymbol{V}_t$ and $\boldsymbol{v}_i(t)$.

5.4.3 Computational Aspects

Let $n = n_{i,i'}$[1] denote the number of basis functions for the hypothesis space $\mathcal{H}_{i,i'}$ when $i \neq i'$ (see the SI for details); when $i = i'$, we simply take $n_{i,i} = 1$. The total computational cost for solving the learning problem is $O(LN^2 + Ld((N-1)n+1)^2 + ((N-1)n+1)^3)$, with $L\frac{N(N-1)}{2} = O(LN^2)$ being spent on computing the pair-wise distances, $O(Ld((N-1)n+1)^2)$ on assembling the linear system for the learning problem, mainly spent on the computation of the matrix-matrix multiplication (see details in SI), $O(((N-1)n+1)^3)$ on solving final linear system of size $((N-1)n+1)^2$. The computational bottleneck comes in the variable L when $L \gg (N-1)n+1$, since we are processing over hundreds of years of data. However, we can parallelize the learning method in L by splitting the long trajectory into pieces, which significantly reduces the computing time and storage, see the SI for details.

The total data needed for the observation of 500 years i.e. $L \approx 500 * 365$) of position/velocity data of $N = 10$ celestial bodies (i.e. $2LNd$) amounts to roughly 11 million data points, hence parallelization in L is needed in order to compute the pairwise quantities efficiently. Meanwhile, during the assembly of the learning matrices, a matrix of size $Ld \times ((N-1)n+1)$ is generated for each celestial body, which requires handling a total of roughly 4.4 billion data points. Hence, the assembly of the learning matrices also has to be done in parallel, see the SI. Once the final matrix is assembled, it is of size $((N-1)n+1) \times ((N-1)n+1)$ (with $n \approx 10^2$), and its inversion can be easily handled.

5.4.4 Modern Ephemerides

We choose the National Aeronautics and Space Administration's (NASA) Jet Propulsion Laboratory's (JPL) Development Ephemeris (DE), numbered as DE430/431, as our source of observation data. These modern ephemerides are routinely updated and

[1] We use a uniform n for all $\mathcal{H}_{i,i'}$ when $i \neq i'$, hence we suppress the dependence of n on (i, i').

maintained, and it has been used in NASA's space exploration missions, and by the Astronomical Almanac, since 1984. Details on DE430/431 may be found in [125].

5.5 Learning Results

We take 500 years of daily position/velocity data $(1500 - 1999)$ of the Sun, 8 major planets, and the Earth's Moon from NASA JPL's DE430/431 from their online database (`https://ssd.jpl.nasa.gov/horizons.cgi`). We perform various learning experiments from subsets of 500 years of daily data, with the acceleration approximated using a Finite Difference Scheme. We present a comparison of our learning results (learned on a training set of 450 years of data with a prediction of an additional 50 years) alongside the observed data, the simulated Newton's system (using the initial positions/velocities at year 1500 from JPL's data), and the simulated Einstein-Infeld-Hoffmann (EIH) system (using the same initial positions/velocities at year 1500), in terms of various trajectory estimation errors, and errors in estimating period/aphelion/perihelion from the trajectories. Additionally, we are interested in how our learned systems can produce the perihelion precession rate of the orbits of Mercury, the Moon, and Mars; we also present the calculated precession rates from JPL's data, Newton, and EIH. Finally, we show a de-coupling procedure to discover the masses directly from the estimators – which is a generalization of the method presented in [146]. For integration on the learned system and simulated Newton's system, we use a symplectic integrator (fourth order Leapfrog with $h = 10^{-2}$); and for the integration on EIH, we use MATLAB's fully implicit integrator, ode15i, with relative tolerance set at 10^{-8} and absolute tolerance at 10^{-11}. See the SI for detailed description of the computer server and computational times spent on various learning/simulation problems.

Before we present the learning results, we describe the setup of our learning trials

in terms of units, constants, and indexing. We adopt the following units to conform to the NASA standard: time unit is $t = 1$ day, length unit is 10^6 km, and the unit of mass is 10^{24} kg. The gravitational constant G and speed of light c have been rescaled in these new units (see SI). We index the celestial bodies as follows: 1 is given to the Sun, 2 to Mercury, $\cdots$, 5 to Earth, 6 to the Moon, 7 to Mars, $\cdots$, lastly 10 to Neptune. For the set of estimators $\widehat{\boldsymbol{\phi}}$, we construct the hypothesis spaces using clamped B-splines, with $(SI, p) = (90, 5)$. Here SI is the number of sub-intervals in each $[R_{i,i'}^{\min}, R_{i,i'}^{\max}]$, and p is the degree of the clamped B-spline functions.

The results are presented in corresponding figures and tables. Figure 5.1 shows the trajectory error estimation for various dynamics compared to the JPL observation data, and to two simulated systems (EIH and Newton). Table 5.1 shows the perihelion precession rates of three prototypical celestial bodies: Mercury, the Moon and Mars estimated from various dynamics: JPL, LD, EIH, and Newton. In order to offer deeper insight into the performance of our learning method, Figure 5.3 shows the errors in estimating the period, aphelion, and perihelion of LD, Newton, EIH when compared to the observation data (JPL), using our own estimation algorithm (see SI).

As shown in Figure 5.3, our learned dynamics excels in almost every estimation, except at estimating the period of Neptune (which might be caused by missing data from Pluto), and estimating the aphelion/perihelion for Mercury, likely due to the relativity effect not being within the collective dynamics modeling framework. Figures 5.4 and 5.5 show the comparison of the estimated interaction kernels (Sun-on-planet and planet-on-Sun interaction kernels) versus Newton's interaction kernels, together with a range of general relativity effects defined by the EIH and projected onto each $\boldsymbol{x}_{i'} - \boldsymbol{x}_i$ range. The error, $\frac{g_{i,i'} - \text{Newton}_{i,i'}}{\text{Newton}_{i,i'}}$, is shown in symmetric-log scale with $\text{Newton}_{i,i'} = \frac{Gm_{i'}}{r^3}$ and $f_{g,i'} = \hat{\phi}_{i,i'}$ or $\text{EIH}_{i,i'}^{\max \text{ or } \min}$. The computation of the EIH range is shown in the SI.

The Sun-on-planet interactions, i.e. $(\hat{\phi}_{i,1})_i$s , drive the movement of the Solar system, due to their massive scale. We are able to recover these kernels in a form between the Newton and EIH level, with point-wise relative errors at the scale around 10^{-5}. For the planet-on-Sun interaction kernels, we are able to de-couple them from the sum of interactions acting on the Sun. These interaction kernels are close to being Newton but improve upon it. Note that the observation data contains the relativity effect, and the results show that our learning method is able to identify the optimal 1D kernel describing the dynamics. In order to understand the perihelion precession rate more deeply, we show in Figure 5.5 the point-wise relative error comparison of the learned interaction kernels vs. Newton and EIH range.

As shown in Figure 5.5, we are able to recover a set of estimators, which oscillate between the maximum or minimum EIH forces around Newton.

5.6 Conclusion

We have demonstrated the effectiveness of our learning methods applied to study celestial mechanics of the solar system using NASA JPL's modern ephemeris. Our learned estimators produce better performance than Newton and EIH in terms of trajectory prediction, and the estimated trajectories also preserve various geometric properties, such as the period, aphelion, perihelion, with high accuracy. Furthermore, our learning methods can produce estimated dynamics which give a perihelion precession rate of Mercury's orbit closer to the observed rate than Newton's law. The estimated dynamics from our $1D$ learning method can also be used to estimate the mass of each celestial body. Aided by geometric machine learning techniques, such as the approach discussed in [93], our learning method can be extended to study galaxy dynamics or the solar system from a relativistic point of view.

5.7 Supplemental Information

5.7.1 Celestial Mechanical Systems

We describe in this section the equations of motion for various celestial mechanical systems: Jet Propulsion Laboratory's system (JPL), Einstein-Infeld-Hoffmann System (EIH), and Newton System.

JPL System

The equations of motion for the JPL system is rather complicated. They consider the contributing celestial bodies (CBs) as follows: Sun, 8 major planets, Moon of Earth, and Pluto. Hence $N = 11$ for the number of CBs in the system. The equations of motion for these CBs are given as follows,

$$
\begin{aligned}
\dot{\boldsymbol{x}}_i &= \boldsymbol{v}_i \\
\dot{\boldsymbol{v}}_i &= \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \frac{G m_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|^3} \cdot (\boldsymbol{x}_{i'} - \boldsymbol{x}_i) \Bigg\{ 1 - \frac{4}{c^2} \sum_{\substack{i''=1 \\ i'' \neq i}}^{N} \frac{G m_{i''}}{\|\boldsymbol{x}_{i''} - \boldsymbol{x}_i\|} - \frac{1}{c^2} \sum_{\substack{i''=1 \\ i'' \neq i'}}^{N} \frac{G m_{i''}}{\|\boldsymbol{x}_{i''} - \boldsymbol{x}_{i'}\|} \\
&\quad + \frac{2\|\boldsymbol{v}_{i'} - \boldsymbol{v}_i\|^2 - \|\boldsymbol{v}_i\|^2}{c^2} - \frac{3}{2c^2} \big(\langle \frac{\boldsymbol{x}_{i'} - \boldsymbol{x}_i}{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|}, \boldsymbol{v}_{i'} \rangle \big)^2 + \frac{1}{2c^2} \langle \boldsymbol{x}_{i'} - \boldsymbol{x}_i, \dot{\boldsymbol{v}}_{i'} \rangle \Bigg\} \\
&\quad - \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \frac{G m_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|^3} \langle \boldsymbol{x}_{i'} - \boldsymbol{x}_i, \frac{3\boldsymbol{v}_{i'} - 4\boldsymbol{v}_i}{c^2} \rangle \cdot (\boldsymbol{v}_{i'} - \boldsymbol{v}_i) \\
&\quad + \frac{7}{2c^2} \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \frac{G m_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|} \dot{\boldsymbol{v}}_{i'} + \sum_{i'=N+1}^{N+3} \frac{G m_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|} (\boldsymbol{x}_{i'} - \boldsymbol{x}_i) + \sum_{297 \text{ asteroids}} \boldsymbol{F},
\end{aligned}
$$

for $i = 1, \ldots, N$. Here G is the gravitational constants, m_i is the mass of the i^{th} CB, c is the speed of light in the vacuum, and $\boldsymbol{x}_i$,$\boldsymbol{v}_i$ represents the position/velocity of the barycenter of the i^{th} CB. The extra 3 CBs are Ceres, Pallas, and Vesta, which are only used for the calculation of $\dot{\boldsymbol{v}}_i$ for the first N CBs. The last term gives the forces from a set of 297 asteroids which are only considered for perturbations on the

Earth, Moon and Mars. The JPL system also considers other possible physical laws, in particular the Lunar Theory, to make the evolution of the celestial motion as close to the true observation data as possible. For details, see [125].

EIH System

The Einstein-Infeld-Hoffmann (EIH) system uses the equations of motion based on a first-order post-Newtonian expansion of Einstein's field equations of general relativity. Given a system of N CBs , indexed by table 5.2, the barycentric acceleration vector of the i^{th} CB is given by

$$
\begin{aligned}
&\dot{\boldsymbol{x}}_i = \boldsymbol{v}_i,\\
&\dot{\boldsymbol{v}}_i = \sum_{\substack{i'=1\\ i'\neq i}}^{N} \frac{Gm_{i'}}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|^3}\cdot(\boldsymbol{x}_{i'}-\boldsymbol{x}_i)\Bigg\{1-\frac{4}{c^2}\sum_{\substack{i''=1\\ i''\neq i}}^{N}\frac{Gm_{i''}}{\|\boldsymbol{x}_{i''}-\boldsymbol{x}_i\|}-\frac{1}{c^2}\sum_{\substack{i''=1\\ i''\neq i'}}^{N}\frac{Gm_{i''}}{\|\boldsymbol{x}_{i''}-\boldsymbol{x}_{i'}\|}\\
&\qquad+\frac{2\|\boldsymbol{v}_{i'}-\boldsymbol{v}_i\|^2-\|\boldsymbol{v}_i\|^2}{c^2}-\frac{3}{2c^2}\big(\langle\frac{\boldsymbol{x}_{i'}-\boldsymbol{x}_i}{\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|},\boldsymbol{v}_{i'}\rangle\big)^2+\frac{1}{2c^2}\langle\boldsymbol{x}_{i'}-\boldsymbol{x}_i,\dot{\boldsymbol{v}}_{i'}\rangle\Bigg\}\\
&\qquad-\sum_{\substack{i'=1\\ i'\neq i}}^{N}\frac{Gm_{i'}}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|^3}\langle\boldsymbol{x}_{i'}-\boldsymbol{x}_i,\frac{3\boldsymbol{v}_{i'}-4\boldsymbol{v}_i}{c^2}\rangle\cdot(\boldsymbol{v}_{i'}-\boldsymbol{v}_i)\\
&\qquad+\frac{7}{2c^2}\sum_{\substack{i'=1\\ i'\neq i}}^{N}\frac{Gm_{i'}}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|}\dot{\boldsymbol{v}}_{i'},
\end{aligned}
$$

for $i = 1, \ldots, N$.

Newton System

The Newton system uses the famous universal law of gravitation as its equations of motion,

$$
\begin{aligned}
&\dot{\boldsymbol{x}}_i = \boldsymbol{v}_i,\\
&\dot{\boldsymbol{v}}_i = \sum_{i'=1,i'\neq i}^{N}\frac{Gm_{i'}}{\|\boldsymbol{x}_{i'}-\boldsymbol{x}_i\|}^3(\boldsymbol{x}_{i'}-\boldsymbol{x}_i), \qquad i=1,\ldots,N.
\end{aligned}
\tag{5.7.1}
$$

5.7.2 Learning Framework

We describe in this section the framework used to find the set of estimators for minimizing the error functional

$$\mathcal{E}_L(\boldsymbol{\varphi}) = \frac{1}{LN}\sum_{l,i=1}^{L,N}\left\|\ddot{\boldsymbol{x}}_i - \sum_{i'=1}^{N}\varphi_{i,i'}(\|\boldsymbol{x}_{i'}(t_l)-\boldsymbol{x}_i(t_l)\|)(\boldsymbol{x}_{i'}(t_l)-\boldsymbol{x}_i(t_l))\right\|^2. \tag{5.7.2}$$

over $\boldsymbol{\varphi} = [\varphi_{i,i'}]_{i,i'=1}^N$ with each $\varphi \in \mathcal{H}_{i,i'}$ and $d(\mathcal{H}_{i,i'}) = n_{i,i'}$.

The loss functional leads to a natural dynamics-adapted probability measure,

$$\rho_{T,i,i'}^L = \frac{1}{L}\sum_{l=1}^{L}\delta_{r_{i,i'}(t_l)}(r), \tag{5.7.3}$$

where $r_{i,i'}(t_l) = \|\boldsymbol{x}_{i'}(t_l)-\boldsymbol{x}_i(t_l)\|$, and δ is understood as a Dirac measure. The measure $\rho_{T,i,i'}^L$ measures the time-averaged appearance of pairwise distance data for estimating the unknown interaction kernel $\phi_{i,i'}$.

Related Works

In [19], a variational approach was introduced for learning the interaction kernel from observations of first order homogeneous particle systems; and convergence properties were analyzed when N, the number of particles, goes to infinity – namely the mean field limit. We extended this learning approach in [89] to heterogeneous particle systems of first and second order; and studied convergence in M, the number of different initial conditions for fixed N. Then in [146], we discussed the steady state behavior of the learned dynamics using the estimated interaction kernels; an extended learning theory on the new second-order models is developed and investigated in [96]. A learning theory on first-order dynamics constrained on Riemannian manifolds is developed and discussed in [93].

Numerical Algorithms

Now we are ready to discuss the algorithm in detail on how to solve (5.7.2) over a set of finite dimensional hypothesis spaces $\{\mathcal{H}_{i,i'}\}_{i,i'=1}^{N}$. First, when $i \neq i'$, we take

$$R_{i,i'}^{\min\,/\,\max} := \{\min\,/\,\max\}_{l=1}^{L} \|\boldsymbol{x}_{i'}(t_l) - \boldsymbol{x}_i(t_l)\|\,;$$

when $i = i'$, we take $R_{i,i}^{\min\,/\,\max} := 0/1$. Next, we use Clamped B-spline functions[2] of degree $p_{i,i'} = p \geq 2$ so that the estimated interaction kernels would at least have continuous first derivatives (we are using a uniform p for all $i \neq i'$) as the basis functions and build $\mathcal{H}_{i,i'}$ over a uniform partition of $[R_{i,i'}^{\min}, R_{i,i'}^{\max}]$ with the number of sub-intervals equaling $S_{i,i'} = S$[3] (note that $n_{i,i'} = n = p + S$ for the Clamped B-spline basis), when $i \neq i'$; when $i = i'$, we simply take $p_{i,i} = 0$ and $S_{i,i} = 1$, hence $n_{i,i} = 1$. Next, for $i = 1, \ldots, N$, we assemble the basis matrices $\Psi_i^l \in \mathbb{R}^{d\times((N-1)n+1)}$, and the right hand size vector $\vec{d}_i^{\,l} \in \mathbb{R}^d$ in the following way. For $\eta_i = 1, \ldots, (N-1)n+1$,

$$\Psi_i^l(:, \eta_i) = \psi_{i,i',\eta_{i,i'}}(r_{i,i'}(t_l))\boldsymbol{r}_{i,i'}(t_l).$$

Here $\eta_{i,i'}$ is computed as follows: when $1 \leq \eta_i \leq n$ (when $i \neq 1$) or 1 (when $i = 1$), we take $\eta_{i,i'} = \eta_i$; when $\eta_i > n$ (when $i > 1$) or 1 (when $i = 1$), we find the i_* such that $\sum_{i''=1}^{i_*} n_{i,i''} < \eta_i < \sum_{i''=1}^{i_*+1} n_{i,i''}$ (recall that $n_{i,i''} = n$ when $i \neq i''$ and $n_{i,i''} = 1$ when $i = i''$), then let $i' = i_* + 1$ and set $\eta_{i,i'} = \eta_i - \sum_{i''=1}^{i_*} n_{i,i''}$. For $\vec{d}_i^{\,l}$, we simply perform the assignment, i.e. set $\vec{d}_i^{\,l} = \dot{\boldsymbol{v}}_i(t_l)$. We then assemble $A_i^L \in \mathbb{R}^{n_i\times n_i}$ and $\vec{b}_i^{\,L} \in \mathbb{R}^{n_i\times 1}$ as follows

$$A_i^L = \frac{1}{L}\sum_{l=1}^{L}(\Psi_i^l)^T\Psi_i^l \quad \text{and} \quad \vec{b}_i^{\,L} = \frac{1}{L}\sum_{l=1}^{L}(\Psi_i^l)^T\vec{d}_i^{\,l}.$$

[2]Other kinds of basis functions can also be used, see examples in [89].

[3]We use a uniform S here to simplify the discussion; in practice, a non-uniform series $S_{i,i'}$s actually helps in reducing the computational time.

We solve for $\widehat{\vec{\alpha}}_i$ from the system $A_i^L\vec{\alpha} = \vec{b}_i^L$ using a pseudoinverse, and assemble $\hat{\phi}_{i,i'} \coloneqq \sum_{\eta_{i,i'}=1}^{n_{i,i'}} \hat{\alpha}_{i,i',\eta_{i,i'}}\psi_{i,i',\eta_{i,i'}}$, here $\psi_{i,i',\eta_{i,i'}}$ is the $\eta_{i,i'}^{th}$ basis function for $\mathcal{H}_{i,i'}^{1D}$. The actual implementation can be easily parallelized in l, see similar implementations done in the SI of [89].

Computational Complexity

We present a detailed discussion on the total computational complexity of solving the learning problem for estimating the unknown interaction kernels. In order to compute the individual learning interval, i.e. $[R_{i,i'}^{\min}, R_{i,i'}^{\max}]$ for $i, i' = 1, \cdots, N$, we need to perform $\frac{N(N-1)}{2}$ computation of pairwise distances at each time instance, hence ending up with a total of $L\frac{N(N-1)}{2} \approx \mathcal{O}(LN^2)$ for computing pairwise distances. Then in assembling Ψ_i^l for each celestial body, the algorithm does $(N-1)n+1$ basis evaluations (it also does $d((N-1)n+1)$ multiplications, but we consider them negligible when compared to function evaluations) at each time step and for each celestial body, thus ending up costing a total of $LN((N-1)n+1)$ basis evaluations. The assembly of $\vec{d}_i^l$ is based on value assignments hence it is negligible. Then for the assembly of A_i^L, it needs to perform a total of $Ld((N-1)n+1)^2$ multiplications when multiplying $(\Psi_i^l)^T\Psi_i^l$ (again we consider addition negligible) for each celestial body, hence we need to do a total of $LNd((N-1)n+1)^2$ operations for assembling A_i^L. Similarly, we need to perform a total of $LNd((N-1)n+1)$ multiplications for the assembly of N $\vec{b}_i^L$s. Finally, in solving the linear systems, it does a total of $N((N-1)n+1)^3$ operations. Therefore, the total computing time needed for the whole learning problem is

$$T_{\text{tol}} = LN^2+LN((N-1)n+1)+LNd((N-1)n+1)^2+LNd((N-1)n+1)+N((N-1)n+1)^3.$$

since we have $Ld \gg ((N-1)n+1)$, i.e. we are processing hundreds of years of observation data, we have $T_{\text{tol}} \approx LNd((N-1)n+1)^2$; the computational bottleneck

is caused by the assembly of the learning matrix A_i^Ls. In the case of learning from 500 years of position/velocity data (i.e. $L \approx 500 * 365$, $N = 10$, $d = 3$, and $n \approx 100$), the assembly of A_i^Ls would require a total of $4.445 \cdot 10^{12}$ operations. Because L is very large (since we have long trajectories), we need to perform the parallelization of learning algorithm in L by splitting the trajectories.

As far as memory is concerned, it takes $2LNd$ to store 500 years of observed position/velocity data of $N = 10$ celestial bodies, amounting to roughly $1.095 \cdot 10^7$ data points, which might still be considered possible by modern day workstations. However, in the assembly of A_i^Ls, one needs to construct all Ψ_i^ls, leading to a total of $LNd((N-1)n+1) \approx 4.933 \cdot 10^9$ data points, pushing into the domain of supercomputers. One remedy would be to process the Ψ_i^ls in sequential order, however turns out to be too slow. The proper choice is to perform the assembly in parallel in L, thus significantly reducing the time needed to perform the assembly in terms of both memory and computing time, and making the learning algorithm able to be run on personal laptops.

Approximating Acceleration Data

In the case of missing acceleration data, i.e., missing $\{\ddot{\boldsymbol{x}}_i(t_l)\}_{i,l=1}^{N,L}$, we will use the 5-point central scheme to approximate the acceleration as follows

$$\dot{\boldsymbol{v}}_i(t_l) \approx \frac{-\boldsymbol{v}_i(t_l+2h) + 8\boldsymbol{v}_i(t_l+h) - 8\boldsymbol{v}_i(t_l-h) + \boldsymbol{v}_i(t_l-2h)}{12h}, \quad h \ll 1$$

Due to the limiting double precision storage of the position/velocity data, this central difference scheme gives the least amount of numerical amplification of noise when $h \geq 10^{-4}$. To show our claim, let us assume that $\boldsymbol{v}_i^\epsilon(t_l) = \boldsymbol{x}_i(t_l) + \boldsymbol{\epsilon}_i^l$, with $\left\|\boldsymbol{\epsilon}_i^l\right\| \approx 10^{-16}$

(i.e. double precision), then

$$\begin{aligned}
&\frac{-\boldsymbol{v}_i^{\epsilon}(t_l+2h)+8\boldsymbol{v}_i^{\epsilon}(t_l+h)-8\boldsymbol{v}_i^{\epsilon}(t_l-h)+\boldsymbol{v}_i^{\epsilon}(t_l-2h)}{12h} \\
&= \frac{-\boldsymbol{v}_i(t_l+2h)+8\boldsymbol{v}_i(t_l+h)-8\boldsymbol{v}_i(t_l-h)+\boldsymbol{v}_i(t_l-2h)}{12h} + \frac{-\boldsymbol{\epsilon}_i^{l+2h}+8\boldsymbol{\epsilon}_i^{l+h}-8\boldsymbol{\epsilon}_i^{l-h}+\boldsymbol{\epsilon}_i^{l-2h}}{12h} \\
&\approx \dot{\boldsymbol{v}}_i(t_l)+\mathcal{O}(h^4)+\frac{3\cdot 10^{-16}}{2h}.
\end{aligned}$$

Hence, the numerical noise gets enhanced by $\frac{3\cdot 10^{-16}}{2h}$. To balance out the need of approximating the acceleration as accurately as possible while not amplifying the numerical noise too much, we need to have

$$h^4 \approx \frac{3\cdot 10^{-16}}{2h} \Rightarrow h^5 \approx \frac{3}{2}\cdot 10^{-16} \Rightarrow h \approx 6.8426\cdot 10^{-4}.$$

Compared to the given $h_{\min} = \frac{1}{24\cdot 60} \approx 6.9444\cdot 10^{-4}$, this 5-point central scheme is our optimal choice, one can show that the 3-point scheme fails to give the desired accuracy with $h_{\min}$ and 7-point (or above) scheme amplifies the numerical noise too much. Similar reasoning would also explain why any central scheme on the position data would not give any comparable accuracy.

Symplectic Integration

In order to preserve the Lagrangian and its associated Hamiltonian system associated to the Newton system, as well as various estimated dynamical systems from our $1D$ estimators, we use a forth order Leapfrog scheme to handle the long time integration and obtain stable enough trajectories. The usual second order Leapfrog scheme tries to integrate the following ODE, $\ddot{\boldsymbol{x}} = F(\boldsymbol{x})$, for $t \in [0,T]$, where $\boldsymbol{x}$ is the position data, with initial conditions $\boldsymbol{x}(0) = \boldsymbol{x}_0$ and $\dot{\boldsymbol{x}}(0) = \boldsymbol{v}(0)$. Then the second order Leapfrog evolves the ODE from t_l to t_{l+1} with the following update scheme

$$\boldsymbol{x}_{t_{l+1}} = \boldsymbol{x}_{t_l} + \boldsymbol{v}_{t_l}\Delta t + \frac{1}{2}F(\boldsymbol{x}_{t_l})\Delta t^2, \quad \boldsymbol{v}_{t_{l+1}} = \boldsymbol{v}_{t_l} + \frac{1}{2}(F(\boldsymbol{x}_{t_l}) + F(\boldsymbol{x}_{t_l+1}))\Delta t.$$

When combined with the Yoshida algorithm, one can obtain a forth order Leapfrog in the following way

$$\begin{aligned}
\boldsymbol{x}_{t_l}^1 &= \boldsymbol{x}_{t_l} + c_1 \boldsymbol{v}_{t_l} \Delta t, & \boldsymbol{v}_{t_l}^1 &= \boldsymbol{v}_{t_l} + d_1 F(\boldsymbol{x}_{t_l}^1) \Delta t, \\
\boldsymbol{x}_{t_l}^2 &= \boldsymbol{x}_{t_l}^1 + c_2 \boldsymbol{v}_{t_l}^1 \Delta t, & \boldsymbol{v}_{t_l}^2 &= \boldsymbol{v}_{t_l}^1 + d_2 F(\boldsymbol{x}_{t_l}^2) \Delta t, \\
\boldsymbol{x}_{t_l}^3 &= \boldsymbol{x}_{t_l}^2 + c_3 \boldsymbol{v}_{t_l}^2 \Delta t, & \boldsymbol{v}_{t_l}^3 &= \boldsymbol{v}_{t_l}^2 + d_3 F(\boldsymbol{x}_{t_l}^3) \Delta t, \\
\boldsymbol{x}_{t_{l+1}} = \boldsymbol{x}_{t_l}^4 &= \boldsymbol{x}_{t_l}^3 + c_4 \boldsymbol{v}_{t_l}^3 \Delta t, & \boldsymbol{v}_{t_{l+1}} &= \boldsymbol{v}_{t_l}^4 = \boldsymbol{v}_{t_l}^3,
\end{aligned}$$

where the (c_1, c_2, c_3, c_4) and (d_1, d_2, d_3) are given as follows

$$\begin{aligned}
w_0 &= -\frac{\sqrt[3]{2}}{2 - \sqrt[3]{2}}, & w_1 &= \frac{1}{2 - \sqrt[3]{2}}, \\
c_1 &= c_4 = \frac{w_1}{2}, & c_2 &= c_2 = \frac{w_0 + w_1}{2}, \\
d_1 &= d_3 = w_1, & d_2 &= w_0.
\end{aligned}$$

Estimating Planet Information

Given only the position data of the celestial bodies, i.e., $\{\boldsymbol{x}_i(t_l)\}_{i,l=1}^{N,L}$ (indexed by table 5.2) for $T_0 = t_1 < \cdots < T_L = T$, observed daily ($t_2 = t_1 = 1$ day), we use the following algorithm to estimate the aphelion, perihelion, period and precession rate of the planets and Moon.

Estimating Masses of Celestial Bodies

Recall the equations of motion which we assume for fitting our learning model to the observation data,

$$\dot{\boldsymbol{v}}_i(t) = \sum_{i'=1}^{N} \phi_{i,i'}(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|)(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)), \quad i = 1, \cdots, N.$$

Algorithm 1 Estimating Planet Information

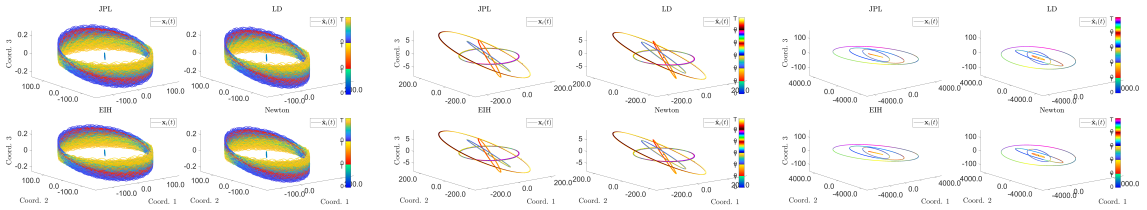
Input: $\{\boldsymbol{x}_i(t_l)\}_{i,l=1}^{N,L}$.
Output: estimated aphelion, perihelion, period, and precession rate.
for $i = 2, \ldots, N$ **do**
    Calculate $\boldsymbol{r}_{1i}(t_l) = \boldsymbol{x}_i(t_l) - \boldsymbol{x}_1(t_l)$.
    Interpolate $br_{1i}(t_l)$ using splines, and obtain $\boldsymbol{r}_{1i}^{\text{spline}}$.
    Evaluate $\boldsymbol{r}_{1i}^{\text{spline}}$ at t_l for $l = 1, \ldots, L'$ with $L' = 24 * 60 * L$.
    Calculate $r_{1i}^{\text{spline}}(t_l) = \left\| \boldsymbol{r}_{1i}^{\text{spline}}(t_l) \right\|$.
    Find local maximum/minimum of the set $\{r_{1i}^{\text{spline}}(t_l)\}_{l=1}^{L'}$.
    From t_a's, where the local maximum of $\{r_{1i}^{\text{spline}}(t_l)\}_{l=1}^{L'}$ are, find the mean/std, which gives the aphelion.
    From t_p's, where the local minimum of $\{r_{1i}^{\text{spline}}(t_l)\}_{l=1}^{L'}$ are, find the mean/std, which gives the perihelion.
    From t_a's and t_p's, find mean/std of the period.
    From t_p's, find out the corresponding position vectors, i.e., $\boldsymbol{r}_{1i}^{\text{spline}}(t_p)$, and calculate the precession advances, θ_p's, from $p = 1$.
    Use the precession model fit, $\theta(t) = \beta_1 + \beta_2 t + \beta_3 t^2$ (according to [104]), then β_2 will give an estimate of the precession rate.

Here the interaction kernel $\phi_{i,i'} : \mathbb{R}^+ \to \mathbb{R}$ is assumed to be either $\phi_{i,i} \equiv 0$ (when $i = i'$) or $\phi_{i,i'}(r) = \frac{U_{i,i'}(r)}{m_{i'}}$ (when $i \neq i'$) for some unknown gravitational potential $U_{i,i'}$, which might depend on m_i and $m_{i'}$ in even a non-linear manner. Although our learning method does not require any knowledge of the masses of the celestial bodies, the estimators obtained from the observation of simple position/velocity data contains hidden information about m_is. In order to gain insights into this additional structure of the estimators, we assume that

$$\hat{\phi}_{i,i'}(r) \approx \beta_{i'} \hat{\phi}_m(r), \quad \text{for } i \neq i', \text{ and } i, i' = 1, \cdots, N.$$

Then we used the non-linear de-coupling optimization based procedure detailed in [146] in order to discover β_is and $\hat{\phi}_m$. Furthermore by assuming that $\hat{\phi}_m(r) = \frac{C}{r^p}$, we found that $p = 2$ gives the best fit, and also discover the masses of the celestial bodies. Results are shown in the main body of this work.

5.7.3 Learning Results



((a)) Earth-Moon-Sun ((b)) Inner Solar ((c)) Outer Solar

Figure 5.6: Comparison of 4 different dynamics: JPL, LD, EIH, and Newton for 3 different types of sub-solar system: Earth-Moon-Sun, inner solar system and outer solar system.

In this section, we present the detailed setup for the learning experiments conducted on 500 years of daily sampled position/velocity data (from year 1500 to 1999) of 10 CBs, and their indices are given in table 5.2.

CB	Sun	Mercury	Venus	Earth	Moon
Index	1	2	3	4	5
CB	Mars	Jupiter	Saturn	Uranus	Neptune
Index	6	7	8	9	10

Table 5.2: Indexing of 10 CBs.

We conform to the units used in NASA's measures, hence we use 10^{24} kg for the unit of mass, 10^6 km for the unit of length, and 1 day for the unit of time, the values for the gravitational constant (G) and speed of light (c) have to be re-scaled, see table 5.3.

G	c
$4.98217402368 \cdot 10^{-4} \frac{(10^6\text{km})^3}{10^{24}\text{kg}\cdot(\text{day})^3}$	$2.59020683712 \cdot 10^4 \frac{10^6\text{km}}{\text{day}}$

Table 5.3: Important Constants

In order to compute the relative errors of the estimation of masses of CBs from our learning algorithm, we use the values for the mass of each CB in table 5.4.

CB	Sun	Mercury	Venus	Earth	Moon
Mass	$1.9885 \cdot 10^6$	0.330	4.87	5.97	0.073
CB	Mars	Jupiter	Saturn	Uranus	Neptune
Mass	0.642	1898	568	86.8	102

Table 5.4: Masses of CBs, unit 10^{24} kg.

We perform various learning experiments on a computing workstation provided by Prisma Analytics, Inc. It has 2 Intel Xeon $E5$-2687w CPUs, each with 12 computing cores, 512 GB memory, and runs on Ubuntu 16.04.7 LTS operating system. The parallel environment is implemented in MATLAB with the "parfor" command. We found that, having implemented the parallelization routine, the computational bottleneck comes from the long-time symplectic integration, which was expected.

Choosing $(SI_{i,i'}, p_{i,i'})$**s**: we conduct various experiments at finding the best $(SI_{i,i'}, p_{i,i'})$ combination (here $SI_{i,i'}$ stands for the number of sub-intervals for $[R_{i,i'}^{\min}, R_{i,i'}^{\max}]$ and $p_{i,i'}$ is the degree of the Clamped B-spline basis for estimating $\phi_{i,i'}$) in terms of trajectory error estimation as well as perihelion precession rate estimation for the Mercury's orbit (we taken the Effective Theory's approach). The combination of $(SI_{i,i'}, p_{i,i'}) = (90, 4)$ gives the best performance.

EIH Range: In order to make the comparison more meaningful, we compare our estimators $\hat{\phi}_{i,i'}$ to Newton and EIH forces for each (i, i') pair. We define

$$\text{Newton}_{i,i'}(r) = \frac{Gm_{i'}}{r^3};$$

and for the EIH range, we first obtain the $\dot{\boldsymbol{v}}_i(t)$ given $\{\boldsymbol{x}_i(t), \boldsymbol{v}_i(t)\}_{i=1}^N$ via (5.7.1), then obtain (over time) the EIH force projected on to $\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)$ as follows

$$\text{EIH}_{i,i'}(t) = \frac{Gm_{i'}}{\|\boldsymbol{x}_{i'} - \boldsymbol{x}_i\|^3} \cdot (\boldsymbol{x}_{i'} - \boldsymbol{x}_i) \Bigg\{ 1 - \frac{4}{c^2} \sum_{\substack{i''=1 \\ i'' \neq i}}^N \frac{Gm_{i''}}{\|\boldsymbol{x}_{i''} - \boldsymbol{x}_i\|} - \frac{1}{c^2} \sum_{\substack{i''=1 \\ i'' \neq i'}}^N \frac{Gm_{i''}}{\|\boldsymbol{x}_{i''} - \boldsymbol{x}_{i'}\|}$$
$$+ \frac{2\|\boldsymbol{v}_{i'} - \boldsymbol{v}_i\|^2 - \|\boldsymbol{v}_i\|^2}{c^2} - \frac{3}{2c^2} \big(\langle \frac{\boldsymbol{x}_{i'} - \boldsymbol{x}_i}{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|}, \boldsymbol{v}_{i'} \rangle \big)^2 + \frac{1}{2c^2} \langle \boldsymbol{x}_{i'} - \boldsymbol{x}_i, \dot{\boldsymbol{v}}_{i'} \rangle \Bigg\},$$

notice that the dependence of $\boldsymbol{x}_i/\boldsymbol{v}_i$ on t is suppressed to simplify the notation. We then compute statistics of $\mathrm{EIH}_{i,i'}$ over $[R^{\min}_{i,i'}, R^{\max}_{i,i'}]$ to obtain the maximum/minimum, i.e. . $\mathrm{EIH}^{\min,\max}_{i,i'}(r)$. Then we consider the following relative errors,

$$\mathrm{Err}^1_{i,i'}(r) = \frac{\mathrm{EIH}^{\max}_{i,i'}(r) - \mathrm{Newton}_{i,i'}(r)}{\mathrm{Newton}_{i,i'}(r)} \quad \text{and} \quad \mathrm{Err}^2_{i,i'}(r) = \frac{\mathrm{EIH}^{\min}_{i,i'}(r) - \mathrm{Newton}_{i,i'}(r)}{\mathrm{Newton}_{i,i'}(r)}$$

and

$$\mathrm{Err}^3_{i,i'}(r) = \frac{\hat{\phi}_{i,i'}(r) - \mathrm{Newton}_{i,i'}(r)}{\mathrm{Newton}_{i,i'}(r)}.$$

Comparison of $\hat{\phi}_{i,i'}$'s: we present the comparisons for $\hat{\phi}_{i,i'}$s for $i = 3, 4, 7, 8, 9, 10$ with the special relative errors defined above, and show them in symmetric log scale so that the details towards zero can be shown properly; it is done via a special transformation which guarantees the continuity across zero [138]. The results are shown in figures 5.7, 5.8, 5.9, 5.10, and 5.11.

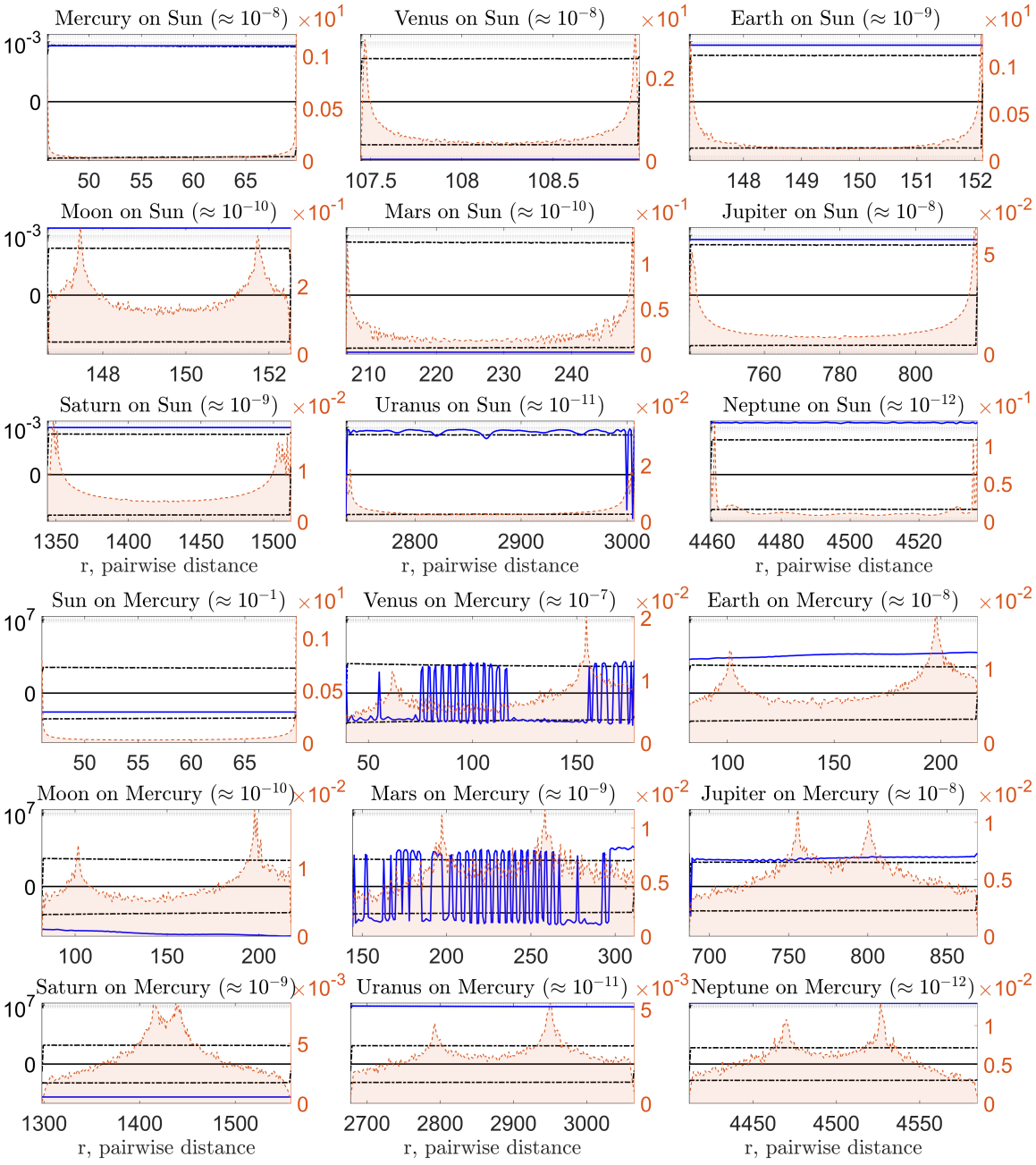


Figure 5.7: CB-on-Sun ($(\hat{\phi}_{1,i})_i$s) and CB-on-Mercury ($(\hat{\phi}_{2,i})_i$s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error compared to Newton in symmetric-log scale. Shown in the background are the corresponding distribution of pairwise distance data used to estimate these kernels, i.e. $\rho^L_{T,i,i'}$s. Dotted black lines represent the errors $\mathrm{Err}^1_{i,i'}(r)$ and $\mathrm{Err}^2_{i,i'}(r)$ for $i = 3, 4$. Solid line black line represents the error for Newton, which is exactly zero. Solid blue line shows the error $\mathrm{Err}^3_{i,i'}(r)$ for $i = 3, 4$. As shown in the sub-plots, our estimators are recovered in a way which is closer to the EIH range then to Newton.

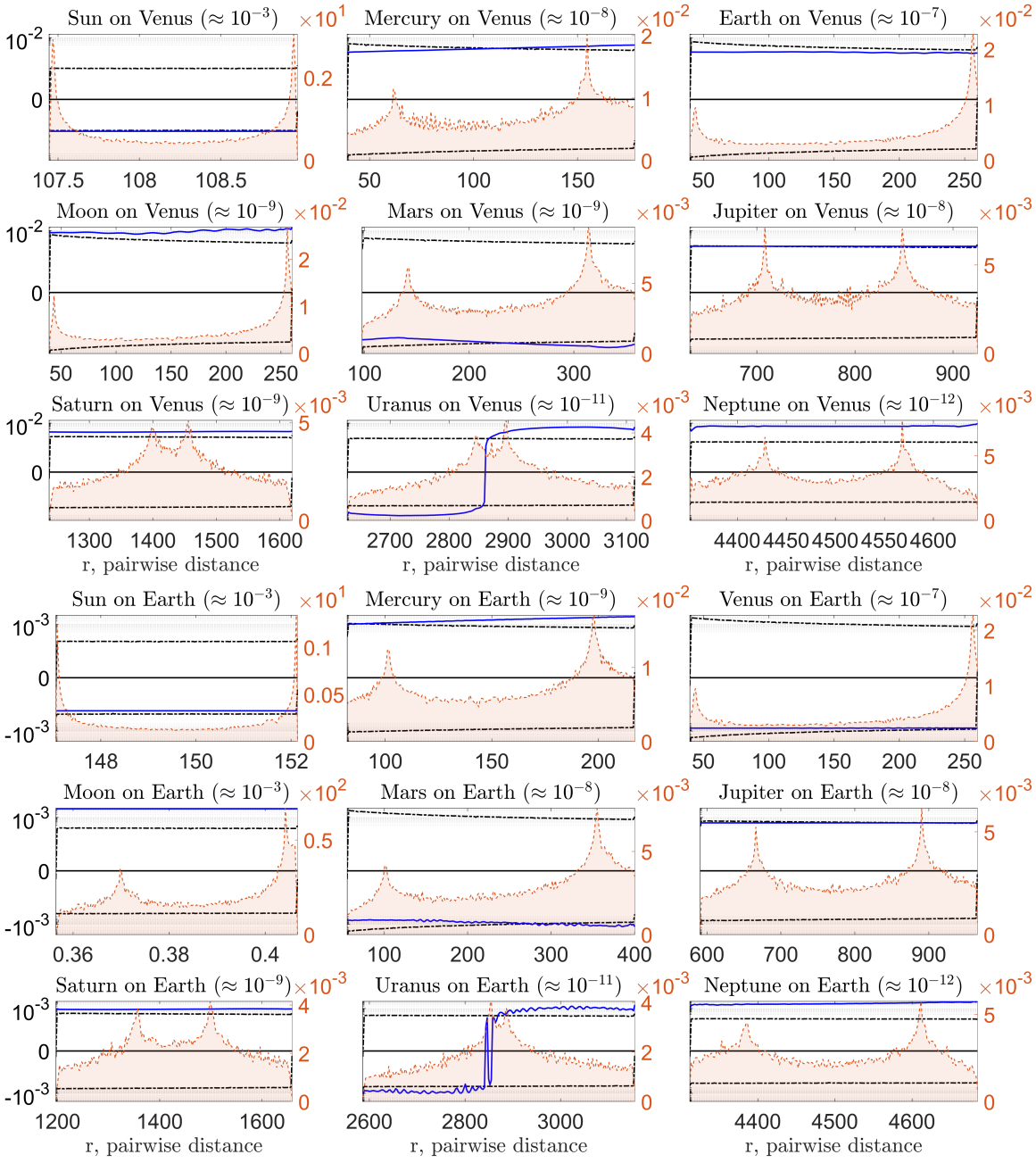


Figure 5.8: CB-on-Venus ($(\hat{\phi}_{3,i})_i$s) and CB-on-Earth ($(\hat{\phi}_{4,i})_i$s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error compared to Newton in symmetric-log scale. Shown in the background are the corresponding distribution of pairwise distance data used to estimate these kernels, i.e. $\rho^L_{T,i,i'}$s. Dotted black lines represent the errors $\mathrm{Err}^1_{i,i'}(r)$ and $\mathrm{Err}^2_{i,i'}(r)$ for $i = 3, 4$. Solid line black line represents the error for Newton, which is exactly zero. Solid blue line shows the error $\mathrm{Err}^3_{i,i'}(r)$ for $i = 3, 4$. As shown in the sub-plots, our estimators are recovered in a way which is closer to the EIH range then to Newton.

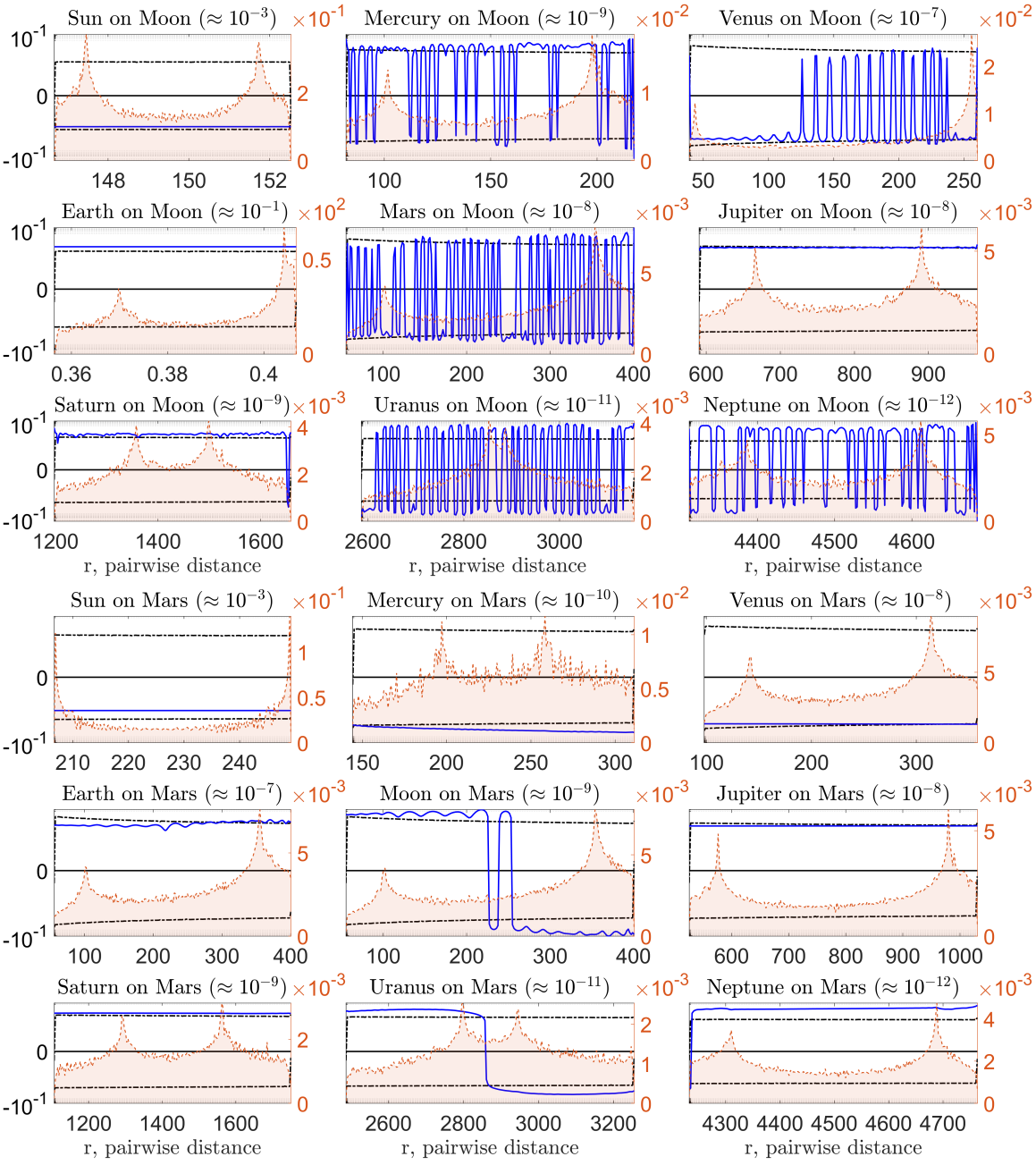


Figure 5.9: CB-on-Moon ($(\hat{\phi}_{5,i})_i$s) and CB-on-Mars ($(\hat{\phi}_{6,i})_i$s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error compared to Newton in symmetric-log scale. Shown in the background are the corresponding distribution of pairwise distance data used to estimate these kernels, i.e. $\rho^L_{T,i,i'}$s. Dotted black lines represent the errors $\mathrm{Err}^1_{i,i'}(r)$ and $\mathrm{Err}^2_{i,i'}(r)$ for $i = 7, 8$. Solid line black line represents the error for Newton, which is exactly zero. Solid blue line shows the error $\mathrm{Err}^3_{i,i'}(r)$ for $i = 7, 8$. As shown in the sub-plots, our estimators are recovered in a way which is closer to the EIH range then to Newton.

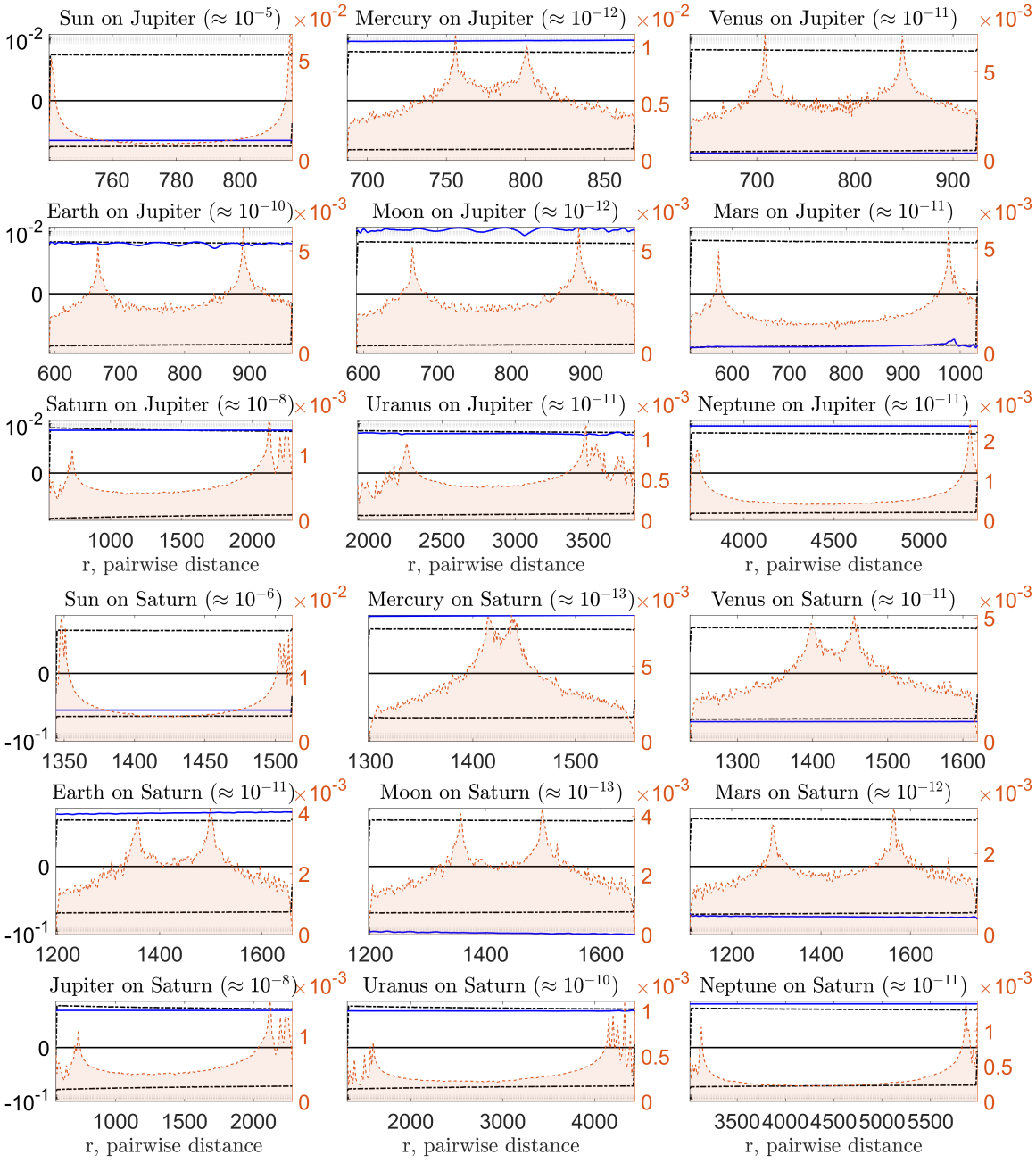


Figure 5.10: CB-on-Jupiter ($(\hat{\phi}_{7,i})_i$s) and CB-on-Saturn ($(\hat{\phi}_{8,i})_i$s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error compared to Newton in symmetric-log scale. Shown in the background are the corresponding distribution of pairwise distance data used to estimate these kernels, i.e. $\rho^L_{T,i,i'}$s. Dotted black lines represent the errors $\text{Err}^1_{i,i'}(r)$ and $\text{Err}^2_{i,i'}(r)$ for $i = 7, 8$. Solid line black line represents the error for Newton, which is exactly zero. Solid blue line shows the error $\text{Err}^3_{i,i'}(r)$ for $i = 7, 8$. As shown in the sub-plots, our estimators are recovered in a way which is closer to the EIH range then to Newton.

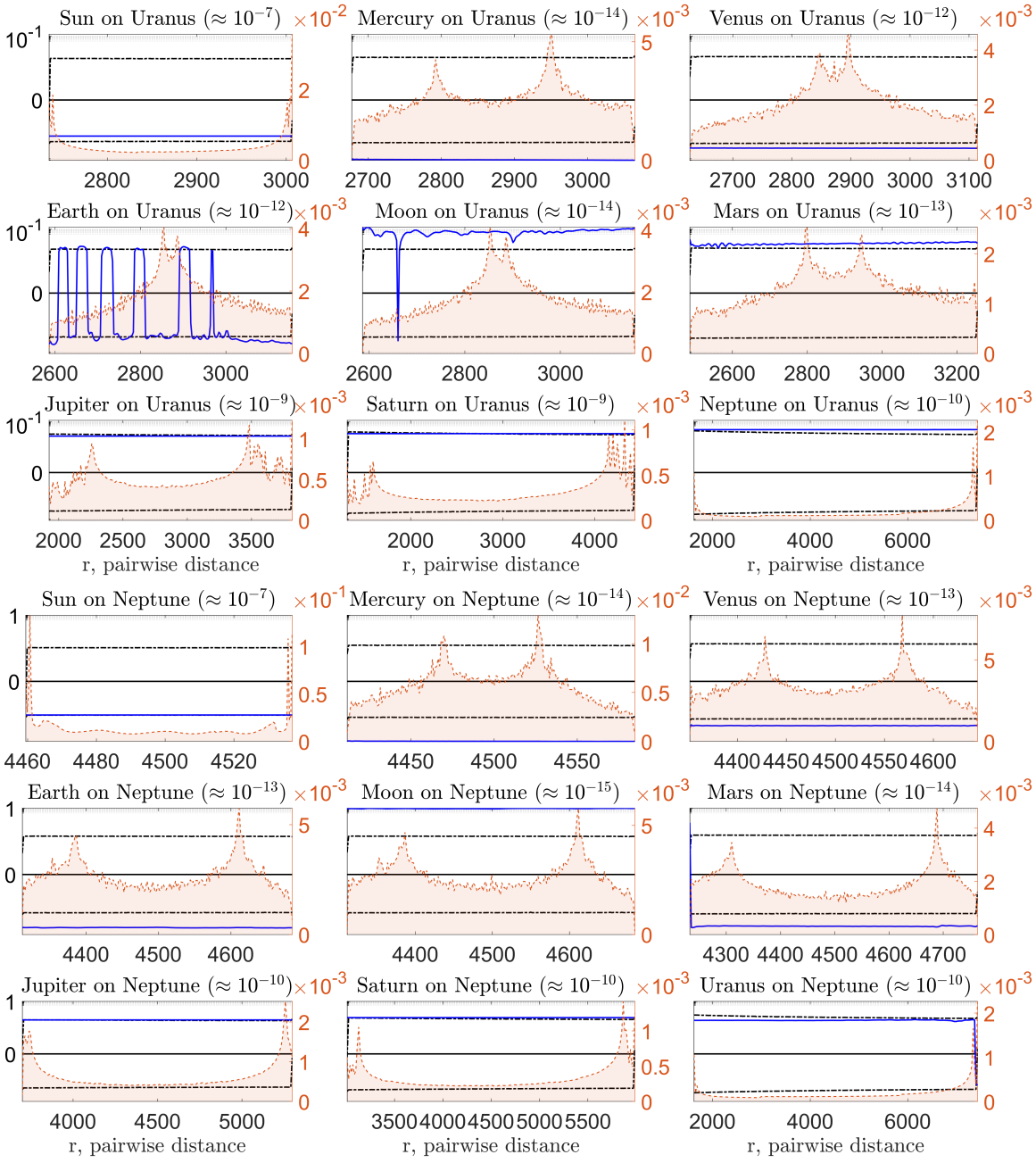


Figure 5.11: CB-on-Uranus ($(\hat{\phi}_{9,i})_i$s) and CB-on-Neptune ($(\hat{\phi}_{10,i})_i$s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error compared to Newton in symmetric-log scale. Shown in the background are the corresponding distribution of pairwise distance data used to estimate these kernels, i.e. $\rho^L_{T,i,i'}$s. Dotted black lines represent the errors $\mathrm{Err}^1_{i,i'}(r)$ and $\mathrm{Err}^2_{i,i'}(r)$ for $i = 9, 10$. Solid line black line represents the error for Newton, which is exactly zero. Solid blue line shows the error $\mathrm{Err}^3_{i,i'}(r)$ for $i = 9, 10$. As shown in the sub-plots, our estimators are recovered in a way which is closer to the EIH range then to Newton.

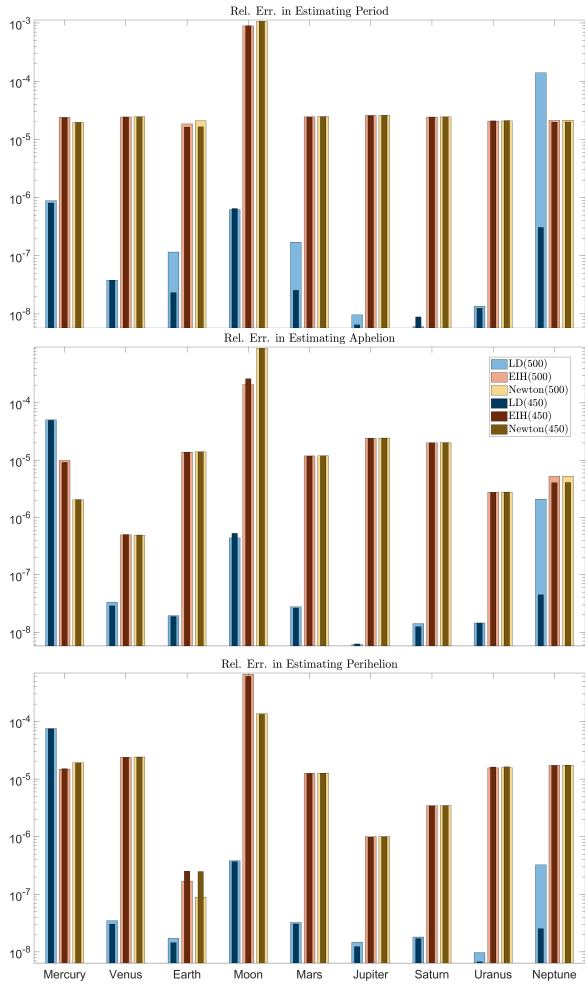


Figure 5.3: Comparison of relative errors in estimating period/aphelion/perihelion from three different dynamics (LD/EIH/Newton) compared to the JPL's observation data for 9 different celestial bodies over 450 and 500 year trajectories. The errors over 450 years have smaller width and darker color, and are laid on top of the errors over 500 years, which have bigger width and lighter color. Different colors correspond to different dynamics: dark/light blue for LD, dark/light red for EIH, and dark/light yellow for Newton. The learned dynamics demonstrates high accuracy in terms of trajectory error in almost all cases.

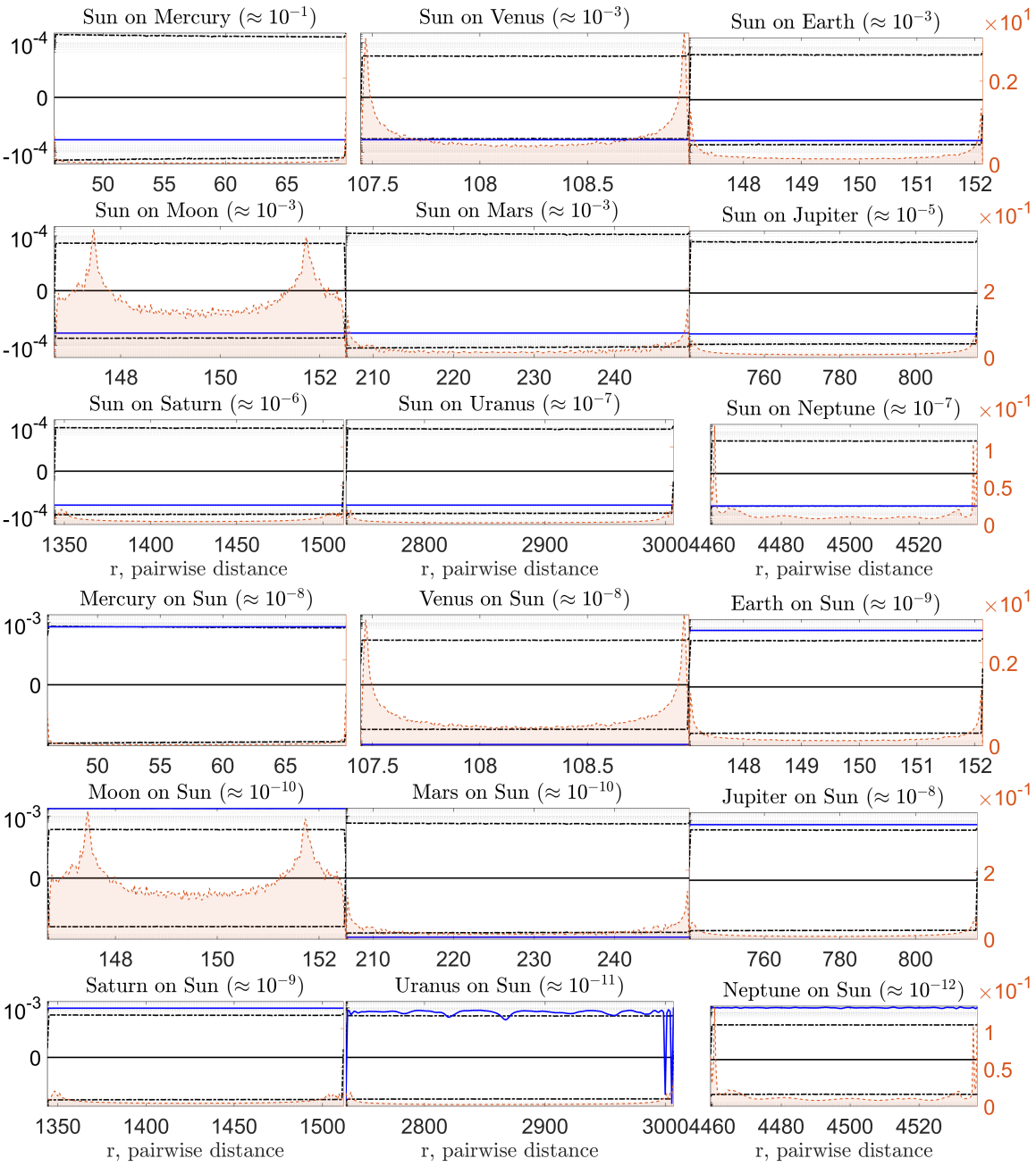


Figure 5.4: Sun-on-planet ($(\hat{\phi}_{i,1})_i$'s) and planet-on-Sun ($(\hat{\phi}_{1,i})_i$'s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error w.r.t Newton in symmetric-log scale. Shown in the background is the corresponding distribution of the pairwise distance data used to estimate these kernels, i.e. $\rho^L_{T,i,i'}$s. We recover the estimators at relative error around 10^{-5} away from Newton's gravity, and within the range of the EIH range for the Sun-on-planet interactions. For the planet-on-Sun interactions, the relative errors are around 10^{-4} away from Newton (especially bigger for the two farthest away planets), due to the fact that the strength of the interactions is getting closer to numerical error of around 10^{-10} coming from the approximated accelerations. The absolute scale of the maximum Newton force for each (i, i')-pair is shown in the title of corresponding sub-plots.

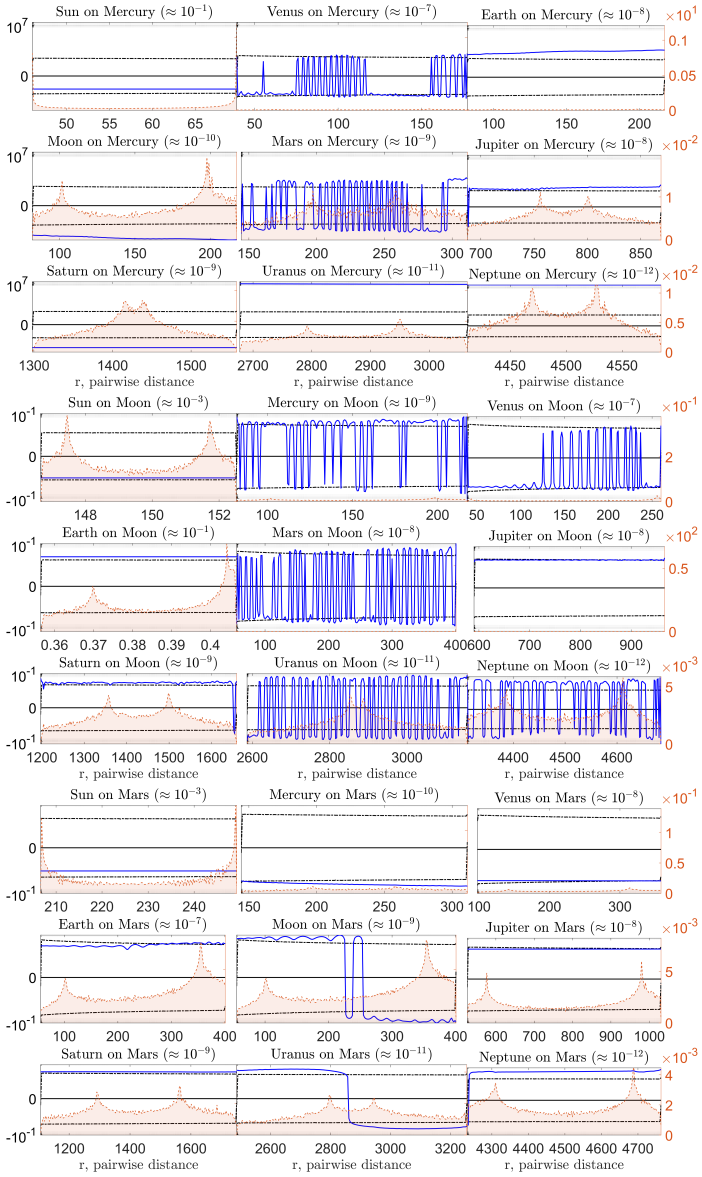


Figure 5.5: Celestial body-on-Mercury ($(\hat{\phi}_{2,i})_i$s), celestial body-on-Moon ($(\hat{\phi}_{5,i})_i$s), and celestial body-on-Mars ($(\hat{\phi}_{6,i})_i$s) interaction kernels vs. Newton and the EIH range, shown in terms of relative error compared to Newton in symmetric-log scale. Shown in the background are the corresponding distribution of pairwise distance data used to estimate these kernels, i.e. $\rho^{L,}_{T,i,i'}$s. As shown in the figures, the learning of interaction kernels on Mars is the easiest; whereas the interaction kernels on the Moon and Mercury both present considerable complications: lunar effects and general relativity effects. Hence the resulting estimators on the Moon and Mercury show more oscillatory behaviors, especially at small scales. The absolute scale of the maximum Newton force for each (i, i')-pair is shown in the title of the corresponding sub-plots.

Bibliography

[1] Nicole Abaid and Maurizio Porfiri. Fish in a ring: Spatio-temporal pattern formation in one-dimensional animal groups. *Journal of the Royal Society Interface*, 7(51):1441–1453, 2010.

[2] B. P. Abbott et al. Tests of general relativity with the binary black hole signals from the ligo-virgo catalog gwtc-1. *Phys. Rev. D*, 100:104036, Nov 2019.

[3] Hyunjin Ahn, Seung-Yeal Ha, Hansol Park, and Woojoo Shim. Emergent behaviors of Cucker-Smale flocks on the hyperboloid. 2020.

[4] Shin Mi Ahn, Heesun Choi, Seung Yeal Ha, and Ho Lee. On collision-avoiding initial configurations to Cucker-smale type flocking models. 10(2):625–643.

[5] G. Albi, D. Balagué, J. A. Carrillo, and J. Von Brecht. Stability analysis of flock and mill rings for second order models in swarming. *SIAM Journal on Applied Mathematics*, 74(3):794–818, 2014.

[6] Aylin Aydoğdu, Sean T. McQuade, and Nastassia Pouradier Duteil. Opinion dynamics on a general compact riemannian manifold. 12:489.

[7] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather

than metric distance: Evidence from a field study. *Proc Natl Acad Sci USA*, 105(4):1232–1237, 2008.

[8] Cosimo Bambi. *Classical Tests of General Relativity*, pages 163–178. Springer Singapore, Singapore, 2018.

[9] Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P. Brenner. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, 2019.

[10] N. Bellomo, P. Degond, and E. Tadmor, editors. *Active Particles, Volume 1*. Springer International Publishing AG, Switerland, 2017.

[11] Andrew J. Bernoff and Chad M. Topaz. A primer of swarm equilibria. 10(1):212–250.

[12] Tom Bertalan, Felix Dietrich, Igor Mezić, and Ioannis G. Kevrekidis. On learning hamiltonian systems from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):121107, 2019.

[13] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

[14] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walzak. Statistical mechanics for natural flocks of birds. *Proc Natl Acad Sci USA*, 109:4786 – 4791, 2012.

[15] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory part i: piecewise constant functions. *Journal of Machine Learning Research*, 6(Sep):1297–1321, 2005.

[16] V. Blodel, J. Hendricks, and J. Tsitsiklis. On Krause's multi-agent consensus model with state-dependent connectivity. *Automatic Control, IEEE Transactions on*, 54(11):2586 – 2597, 2009.

[17] J. Bongard and H. Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 104(24):9943–9948, 2007.

[18] Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.

[19] M. Bongini, M. Fornasier, M. Hansen, and M. Maggioni. Inferring interaction rules from observations of evolutive systems I: The variational approach. *Math Mod Methods Appl Sci*, 27(05):909–951, 2017.

[20] C. Brugna and G. Toscani. Kinetic models of opinion formation in the presence of personal conviction. *Physical Review E*, 92(5):052818, 2015.

[21] N. Brunel. Parameter estimation of ODEs via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.

[22] S. Brunton, N. Kutz, and J. Proctor. Data-drive discovery of governing physical laws. *SIAM News*, 50(1), 2017.

[23] S. Brunton, J. Proctor, and J. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):3932–3937, 2016.

[24] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[25] J. Cao, L. Wang, and J. Xu. Robust estimation for ordinary differential equation models. *Biometrics*, 67(4):1305–1313, 2011.

[26] Marco Caponigro, Anna Lai, and Benedetto Piccoli. A nonlinear model of opinion formation on the sphere. 35.

[27] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91:045002, Dec 2019.

[28] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[29] J. Carrillo, M. D'Orsogna, and V. Panferov. Double Milling in self-propelled swarms from kinetic theory. *Kinetic & Related Models*, 2(2):363 – 378, 2009.

[30] Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019.

[31] Minshuo Chen, Hao Liu, Wenjing Liao, and Tuo Zhao. Doubly robust off-policy learning on low-dimensional manifolds by deep neural networks.

[32] Y. Chen and T. Kolokolnikov. A minimal model of predator-swarm interactions. *J. R. Soc. Interface*, 11:20131208, 2013.

[33] Hillel J. Chiel, Jeffrey P. Gill, Jeffrey M. McManus, and Kendrick M. Shaw. Learning biology by recreating and extending mathematical models. *Science*, 336(6084):993–994, 2012.

[34] Yeol Cho, S. Sever, and Young-Ho Kim. On some Gronwall type inequalities with iterated integrals. *Mathematical Communications*, 12(1):63–73, 2007.

[35] Young Pil Choi, Seung Yeal Ha, and Zhuchun Li. Emergent dynamics of the cucker–Smale flocking model and its variants. (9783319499949):299–331.

[36] Y. Chuang, Y. Huang, M. D'Orsogna, and A. Bertozzi. Multi-vehicle flocking: scalability of cooperative control algorithms using pairwise potentials. *IEEE International Conference on Robotics and Automation*, pages 2292 – 2299, 2007.

[37] Yao-Li Chuang, Tom Chou, and Maria R. D'Orsogna. Swarming in viscous fluids: Three-dimensional patterns in swimmer- and force-induced flows. *Physical Review E*, 93(4):1–12, 2016.

[38] Yao-li Chuang, Maria R. D'Orsogna, Daniel Marthaler, Andrea L. Bertozzi, and Lincoln S. Chayes. State transitions and the continuum limit for a 2D interacting, self-propelled particle system. *Physica D: Nonlinear Phenomena*, 232(1):33–47, 2007.

[39] Antonio C. Costa, Tosif Ahamed, and Greg J. Stephens. Adaptive, locally linear models of complex dynamics. *Proceedings of the National Academy of Sciences*, 116(5):1501–1510, 2019.

[40] I. Couzin, J. Krause, N. Franks, and S. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513 – 516, 2005.

[41] F. Cucker and J.-G. Dong. A general collision-avoiding flocking framework. *IEEE Trans. Automat. Control*, 56(5):1124 – 1129, 2011.

[42] F. Cucker and J.-G. Dong. A conditional, collision-avoiding, model for swarming. *Discrete and Continuous Dynamical Systems*, 43(3):1009 – 1020, 2014.

[43] F. Cucker and E. Mordecki. Flocking in noisy environments. *J. Math. Pures Appl.*, 89(3):278 – 296, 2008.

[44] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc*, 39(1):1–49, 2002.

[45] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

[46] F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Transactions on automatic control*, 52(5):852, 2007.

[47] F. Cucker and S. Smale. On the mathematics of emergence. *Jpn. J. Math.*, 2(1):197 – 227, 2007.

[48] Felipe Cucker and Jiu Gang Dong. Avoiding collisions in flocks. 55(5):1238–1243.

[49] Felipe Cucker and Steve Smale. Emergent behavior in flocks. 52(5):852–862.

[50] T. Cui, Y. Marzouk, and K. Willcox. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966 – 990, 2014.

[51] W. Dahmen, R. DeVore, and K. Scherer. Multi-Dimensional Spline Approximation. *SIAM J. Numer. Anal.*, 17(3):380–402, 1980.

[52] I. Dattner and C. Klaassen. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9(2):1939–1973, 2015.

[53] C. de Boor and R. DeVore. Approximation by Smooth Multivariate Splines. *Transactions of the American Mathematical Society*, 276(2):775, 1983.

[54] Pierre Degond, Jian-Guo Liu, Sebastien Motsch, and Vladislav Panferov. Hydrodynamic models of self-organized dynamics: Derivation and existence theory. 20(2):89–114.

[55] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Approximation methods for supervised learning. *Foundations of Computational Mathematics*, 6(1):3–58, 2006.

[56] Manfredo P. do Carmo. *Diferential Geometry of Curves and Surfaces*. Prentice-Hall.

[57] Marina Dolfin and Miroslaw Lachowicz. Modeling opinion dynamics: How the network enhances consensus. 10(4):877–896.

[58] Carmeline J. Dsilva, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. *Applied and Computational Harmonic Analysis*, 44(3):759 – 773, 2018.

[59] Radek Erban, Jan Haškovec, and Yongzheng Sun. A Cucker-Smale model with noise and delay. 76(4):1535–1557.

[60] Razvan Fetecau and Beril Zhang. Self-organization on riemannian manifolds. 11(3):397–426.

[61] Veysel Gazi. On lagrangian dynamics based modeling of swarm behavior. 260:159 – 175. Emergent Behaviour in Multi-particle Systems with Non-local Interactions.

[62] G. Grégoire and H. Chaté. Onset of collective and cohesive motion. *Phy. Rev. Lett.*, 92, 2004.

[63] Anupam Gupta, Amal Roy, Arnab Saha, and Samriddhi Sankar Ray. Flocking of Active Particles in a Turbulent Flow. pages 1–6.

[64] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, New York, 2002.

[65] Seung Yeal Ha and Jian Guo Liu. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. 7(2):297–325.

[66] E. Hairer. Geometric integration of ordinary differential equations on manifolds. 41(5):996 – 1007.

[67] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer.

[68] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year enso forecasts. *Nature*, 573(7775):568–572, 2019.

[69] Jiequn Han, Chao Ma, Zheng Ma, and Weinan E. Uniformly accurate machine learning-based hydrodynamic models for kinetic equations. *Proceedings of the National Academy of Sciences*, 116(44):21983–21991, 2019.

[70] X. Han, Z. Shen, W. Wang, and Z. Di. Robust reconstruction of complex networks from sparse data. *Physical Review Letters*, 114(2):028701, 2015.

[71] Pierre-Emmanuel Jabin and Sebastien Motsch. Clustering and asymptotic behavior in opinion formation. 257(11):4165 – 4187.

[72] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[73] S. Kang, W. Liao, and Y. Liu. Ident: Identifying differential equations with numerical time evolution. *arXiv preprint arXiv:1904.03538*, 2019.

[74] Y. Katz, K. Tunstrom, C. Ioannou, C. Huepe, and I. Couzin. Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences of the United States of America*, 108:18720–8725, 2011.

[75] Rachael Keller and Qiang Du. Discovery of dynamics using linear multistep methods.

[76] U. Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.

[77] Nikita Kruk, Yuri Maistrenko, and Heinz Koeppl. Self-propelled chimeras. 98(3):1–15.

[78] Y. Kuramoto. Lecture notes in physics. In *Internatonal Symposium on Mathematical Problems in Theoretical Physics*, page 420. Springer-Verlag.

[79] John M. Lee. *Introduction to Smooth Manifolds*. Springer.

[80] Seungjoon Lee, Mahdi Kooshkbaghi, Konstantinos Spiliotis, Constantinos I. Siettos, and Ioannis G. Kevrekidis. Coarse-scale pdes from fine-scale observations via machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1):013141, 2020.

[81] T. Lee, M. Leok, and N. H. McClamroch. *Global Formations of Lagrangian and Hamiltonian Dynamics on Manifolds: A Geometric Approach to Modeling and Analysis*. Springer.

[82] Demian Levis, Albert Diaz-Guilera, Ignacio Pagonabarraga, and Michele Starnini. Flocking and spreading dynamics in populations of self-propelled agents. pages 1–12.

[83] Qianxiao Li, Felix Dietrich, Erik M. Bollt, and Ioannis G. Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):103111, 2017.

[84] Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, and Cheng Zhang. On the identifiability of interaction functions in systems of interacting particles. *arXiv preprint arXiv:1912.11965*, 2019.

[85] H. Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

[86] Thomas S. Logsdon. *Orbital mechanics: theory and applications*. John Wiley, 1998.

[87] Z. Long, Y. Lu, X. Ma, and B. Dong. PDE-net: Learning PDEs from data. *arXiv preprint arXiv:1710.09668*, 2017.

[88] Malcolm Longair. *Theoretical Concepts in Physics*. Cambridge University Press, 3rd edition, 2020.

[89] F. Lu, M. Zhong, S. Tang, and M. Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences of the United States of America*, 116(29):14424–14433, 2019.

[90] Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories, 2019.

[91] R. Lukeman, Y. Li, and L. Edelstein-Keshet. Inferring individual rules from collective behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 107:12576 – 12580, 2010.

[92] Ryan Lukeman, Yue-Xian Li, and Leah Edelstein-Keshet. A conceptual model for milling formations in biological aggregates. 71(2):352.

[93] M. Maggioni, J. Miller, H. Qiu, and M. Zhong. Learning interaction kernels for agent systems on riemannian manifolds, 2021.

[94] Mauro Maggioni, Jason Miller, and Ming Zhong. Agent-based learning of celestial dynamics from ephemerides. *In preparation*, 2020.

[95] H. Miao, X. Xia, A. Perelson, and H. Wu. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review*, 53(1):3–39, 2011.

[96] Jason Miller, Sui Tang, Ming Zhong, and Mauro Maggioni. Learning theory for inferring interaction kernels in second-order interacting agent systems.

[97] Helio H. L. C. Monte-Alto, Mariela Morveli-Espinoza, and Cesar A. Tacla. Multi-agent systems based on contextual defeasible logic considering focus.

[98] S. Mostch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577 – 621, 2014.

[99] H. Niwa. Self-organizing dynamic model of fish schooling. *J. Theor. Biol.*, 171:123 – 136, 1994.

[100] Kevin O'Keeffe and Christian Bettstetter. A review of swarmalators and their potential in bio-inspired computing. page 85, 2019.

[101] Kevin P. O'Keeffe, Joep H.M. Evers, and Theodore Kolokolnikov. Ring states in swarmalator systems. *Physical Review E*, 98(2), 2018.

[102] Kevin P. O'Keeffe, Hyunsuk Hong, and Steven H. Strogatz. Oscillators that sync and swarm. *Nature Communications*, 8(1):1–12, 2017.

[103] Randal Olson, Arend Hintze, Fred Dyer, Jason Moore, and Christoph Adami. Exploring the coevolution of predator and prey morphology and behavior.

[104] Ryan S. Park, William M. Folkner, Alexander S. Konopliv, James G. Williams, David E. Smith, and Maria T. Zuber. Precession of Mercury's Perihelion from Ranging to the MESSENGER Spacecraft . *The Astronomical Journal*, 153(3):121, 2017.

[105] Lydia Patton. Expanding theory testing in general relativity: Ligo and parametrized theories. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 69:142 – 153, 2020.

[106] H. Qin. Machine learning and serving of discrete field theories. *Sci. Rep.*, 10:19329, Nov 2020.

[107] M. Raissi. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *The Journal of Machine Learning Research*, 19(1):932–955, 2018.

[108] M. Raissi and G. Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357:125–141, 2018.

[109] M. Raissi, P. Perdikaris, and G. Karniadakis. Multistep neural networks for data-driven discovery of nonlinear dynamical systems. *arXiv preprint arXiv:1801.01236*, 2018.

[110] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020.

[111] J. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

[112] Francesco Riccio, Roberto Capobianco, and Daniele Nardi. DOP: deep optimistic planning with approximate value function evaluation. abs/1803.08501.

[113] A E Roy. *Orbital Motion*. Taylor and Francis Group, 4th edition, 2005.

[114] H. Rudy, N. Kutz, and S. Brunton. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *Journal of Computational Physics*, 2019.

[115] S. Rudy, S. Brunton, J. Proctor, and N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

[116] Lars Ruthotto, Stanley J. Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 2020.

[117] Alain Sarlette and Rodolphe Sepulchre. Consensus optimization on manifolds. 48.

[118] H. Schaeffer, R. Caflisch, C. Hauck, and S. Osher. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17):6634–6639, 2013.

[119] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[120] Larry Schumaker. *Spline Functions: Basic Theory*. Cambridge University Press, 3rd edition, 2007.

[121] Ruiwen Shu and Eitan Tadmor. Anticipation breeds alignment. *arXiv preprint arXiv:1905.00633*, 2019.

[122] Ruiwen Shu and Eitan Tadmor. Flocking hydrodynamics with external potentials. *Arch Rational Mech Anal*, (238):347 – 381, 2020.

[123] Roman Shvydkoy and Eitan Tadmor. Eulerian dynamics with a commutator forcing. 1(1):1–26.

[124] Christian Soize and Roger Ghanem. Probabilistic learning on manifolds constrained by nonlinear partial differential equations for small datasets.

[125] E. Standish and J. Williams. Chapter 8 : Orbital ephemerides of the sun , moon , and planets. 2007.

[126] S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, 1986.

[127] S. H. Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D*, 143(143):1 – 20, 2000.

[128] K. Tonstrom, Y. Katz, C. C. Ioannou, C. Huepe, M. J. Kutz, and I. D. Couzin. Collective states, multistability and transitional behavior in schooling fish. *Computational Biology*, 9, February 2013.

[129] G. Tran and R. Ward. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling and Simulation*, 15(3):1108–1129, 2017.

[130] G. Tran and R. Ward. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, 15(3):1108 – 1129, 2017.

[131] J. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[132] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[133] Slava Turyshev. Experimental tests of general relativity. *Annual Review of Nuclear and Particle Science*, 58:207–248, 2008.

[134] Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer Publishing Company, Incorporated, 1st edition, 1996.

[135] J. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, 1982.

[136] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel Type of Phase Transition in a System of Self-Driven Particles. *Physical Review Letters*, 75:1226–1229, August 1995.

[137] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel Type of Phase Transition in a System of Self-Driven Particles. *Physical Review Letters*, 75(6):1226 – 1229, 1995.

[138] J. B. Webber. A bi-symmetric log transformatoin for wide-range data . *Measurement Science and Technology*, 24(2):027001, 2012.

[139] G. Weisbuch, G. Deffuant, F. Amblard, and J.-P. Nadal. Interacting agents and continuous opinions dynamics. In Robin Cowan and Nicolas Jonard, editors, *Heterogenous Agents, Interactions and Economic Performance*, pages 225–242. Springer Berlin Heidelberg.

[140] James D. Wells. *Effective Theories in Physics*. Springer-Verglad Berlin Heidelberg, 1st edition, 2012.

[141] Krzysztof Wróbel, Pawel Torba, Maciej Paszyński, and Aleksander Byrski. Evolutionary multi-agent computing in inverse problems. 14.

[142] Or Yair, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Reconstruction of normal forms by learning informed observation geometries from data. *Proceedings of the National Academy of Sciences*, 114(38):E7865–E7874, 2017.

[143] Or Yair, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Reconstruction of normal forms by learning informed observation geometries from data. *Proceedings of the National Academy of Sciences*, 114(38):E7865–E7874, 2017.

[144] Shihao Yang, Samuel W. K. Wong, and S. C. Kou. Inference of dynamic systems from noisy and sparse data via manifold-constrained gaussian processes.

[145] S. Zhang and G. Lin. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217):20180305, 2018.

[146] Ming Zhong, Jason Miller, and Mauro Maggioni. Data-driven discovery of emergent behaviors in collective dynamics. *Physica D: Nonlinear Phenomena*, page 132542, 2020.

Vita

Jason Miller was born in Rockville Maryland, but grew up in Northern Virginia. He received his bachelors of arts degree and master of arts degree, both in mathematics, from the University of Virginia. After graduation, he worked as a quantitative research analyst before deciding to return to academia to pursue a Ph.D. in applied mathematics and statistics (AMS) at Johns Hopkins University. He completed a Masters of Science in Engineering in AMS in 2019 and will complete his Ph.D. in the summer of 2021 after a wonderful 4 years at Johns Hopkins. Jason has been very fortunate to be advised by Mauro Maggioni and has worked with collaborators and mentors from the departments of applied mathematics and statistics, mathematics, as well as the Johns Hopkins School of Medicine. His research interests are broadly in machine learning, nonparametric statistics, the intersection of physics and machine learning, and the intersection of medicine and machine learning. While at Johns Hopkins, Jason has received an Institute for Computational Medicine Fellowship for his research on novel machine learning methods for medical data, is a teaching fellow, and won the Joel Dean Award for Excellence in Teaching.